DATA-DRIVEN APPROACHES FOR PARAPHRASING ACROSS LANGUAGE VARIATIONS

by

Wei Xu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science
New York University
January, 2014

Professor Ralph Grishman

Dedication

To my cat Pupu.

Acknowledgments

Studying for a PhD and moving to a foreign country has been a valuable and exciting experience. I could not have succeeded without the advice, support and influence of many colleagues, friends and family. My thanks are due first and foremost to my advisor, Ralph Grishman, for his guidance and continuous support throughout my graduate career. Ralph gave me consistently good advice and challenges, the opportunity to present at government research meetings and work with people from many different institutes, and tremendous freedom to explore my industrial and research interests. His professional expertise with a sense of humor and positive attitude will always been an inspiration to me. I will be forever grateful for his help.

During my PhD, I have been immensely fortunate to work and co-author with many great individuals. I was lucky to have the opportunity to intern at ETS, Microsoft and Amazon.com. My mentors Joel Tetreault and Martin Chodorow gave me the confidence and tons of help to work out difficult situations. Bill Dolan always provided different and more significant angles of thinking. My manager Xiaodong Fan demonstrated great leadership in both technical and management sides. I also have had the fortune to visit University of Washington in the later part of my graduate studies and benefit greatly from interactions with people over there, Mausam, Oren Etzioni, Luke Zettlemoyer, Fei

Xia, Emily Bender, Alan Ritter, Ralphael Hoffmann, Yoav Artzi, Mark Yatskar, Thomas Lin, Jeff Huang, Nicholas FitzGerald. I am also grateful to the openness and generosity of Kathy McKeown and Heng Ji, who made it possible for me to attend many interesting talks and workshops at Columbia Univeristy and City University of New York. I also received great help from Le Zhao, who was a student at Carnegie Mellon University at the time, with his expertise in the Information Retrieval field.

I owe special thanks to many faculty and students at New York University. Satoshi Sekine and Adam Meyers always gave me good suggestions and generous help. Ernie Davis introduced me to psycholinguistics. I would like to thank Allan Gottlieb, Jingyang Li, Eric Hielscher, Arthur Meacham, Sunandan Chakraborty, Aditya Dhananjay, Nektarios Paisios and members of the Proteus group - Cristina Mota, Shasha Liao, Ang Sun, Bonan Min, Masha Pershina and Shoji Fujiwara. I would also like to thank Leslie Cerve and Rosemary Amico who went out their way and made my graduate study smooth.

Abstract

Our language changes very rapidly, accompanying political, social and cultural trends, as well as the evolution of science and technology. The Internet, especially the social media, has accelerated this process of change. This poses a severe challenge for both human beings and natural language processing (NLP) systems, which usually only model a snapshot of language presented in the form of text corpora within a certain domain and time frame.

While much previous effort has investigated monolingual paraphrase and bilingual translation, we focus on modeling meaning-preserving transformations between variants of a single language. We use Shakespearean and Internet language as examples to investigate various aspects of this new paraphrase problem, including acquisition, generation, detection and evaluation.

A data-driven methodology is applied intensively throughout the course of this study. Several paraphrase corpora are constructed using automatic techniques, experts and crowdsourcing platforms. Paraphrase systems are trained and evaluated by using these data as a cornerstone. We show that even with a very noisy or a relatively small amount of parallel training data, it is possible to learn paraphrase models which capture linguistic phenomena. This work expands the scope of paraphrase studies to targeting

different language variations, and more potential applications, such as text normalization and domain adaptation.

Table of contents

TABLE OF CONTENTS

List of Figures

List of Figures

List of Tables

List of Tables

Chapter 1

Introduction

Our language is a living language, and as such, it is constantly changing. Since William Shakespeare coined hundreds if not thousands of words 450 years ago (Elliott and Valenza, 2011), the English language has evolved greatly under the influence of new communication technologies. The recent addition of 2,000 new entries, including Internet slang, to the authoritative Oxford English Dictionary is a clear example of this process. In addition, varieties of language are warranted by different settings in which they are used or different people who are using them. Unfortunately most natural language processing (NLP) technologies are developed in the context of static corpora, without taking varieties of language into account (Mota and Grishman, 2009, 2008). The performance of standard NLP tools is severely degraded when used for processing Early Modern English or Internet English (Foster et al., 2011; Gimpel et al., 2011; Liu et al., 2011d; Ritter et al., 2011b). Many domain adaptation methods have been proposed; however, the linguistic differences in evolving languages are greater and more structural than these methods normally face. Paraphrase models which track evolving

language could directly benefit a wide range of language technologies (including Information Retrieval, Information Extraction, Summarization and Question Answering), by automatically adapting them to handle the language variations.

For example, we can paraphrase the inputs into a style that is familiar to the system (e.g. contemporary Standard English, as shown in Table 1.1), or alternatively transfer linguistic annotations from Standard English to historic or social media text using aligned parallel corpora. In order to understand natural language, systems must be capable of recognizing synonymous expressions across language varieties, such as spelling variations, syntactic variants, synonyms of words and phrases, and more complex sentence-level variations.

Shakespearean	**Standard English**
Romeo slew Tybalt.	Romeo killed Tybalt.
Tybalt, the kinsman to old Capulet, hath sent a letter to his father's house.	Tybalt, old Capulet's relative, has sent a letter to his father's house.
Internet English	**Standard English**
Hostess is going outta biz	Hostess is going out of business.
UMG COO Voted Off The Island	UMG COO stepped down.
YOU CANT OVERSTATE JUST HOW SIGNIFICANT #EGYPT PM ' S VISIT TO #GAZA IS	You can't overstate just how significant the Egyptian Prime Minister's visit to Gaza is.

Table 1.1: Semantically equivalent expressions

On the other hand, systems capable of paraphrasing text targeting a specific linguistic and writing style could be useful for a variety of applications. They could:

1. Benefit educational applications, allowing students to:

    a) Access *modern English* versions of historical works.

   b) Experiment with writing in the style of an author they are studying.

2. Help marketers to tailor messaging for different social media platforms.

3. Enable non-experts to better consume technical information, for example by translating legalese or medical jargon into nontechnical English.

While there has been much previous effort studying monolingual paraphrase (McKeown, 1979) and bilingual translation (Brown et al., 1993), ***this thesis introduces the task of paraphrasing across varieties of a single language and presents the first methodologies for collecting datasets, building and evaluating models for this task***.

1.1 What is a Paraphrase?

Paraphrases are different words, phrases or sentences that express the same or almost the same meaning. For example, "forget" is a paraphrase of "fail to remember". The criteria of semantically equivalence —"the same or almost the same meaning" —are difficult to define exactly and can vary from task to task. For example, whether "car" is a proper paraphrase of "vehicle" depends on the context and the purpose of paraphrasing.

As discussed by Madnani and Dorr (2010), paraphrases can be divided into 3 categories: Lexical, Phrasal and Sentential. Lexical paraphrases refer to words whose meaning is nearly equivalent in context. Examples include synonyms (e.g. "handgun" vs. "pistol"), hypernyms ("rifle" vs. "firearm") and meronyms ("Chicago" vs. "Illinois"). Phrasal paraphrases refer to phrases which can have equivalent meaning in context ("want to see", "would be nice to visit"), ("lived in Chicago", "grew up in Illinois"). Finally sentential paraphrases are entire sentences which express the same meaning. The criteria

for deciding which words, phrases and sentences are considered as paraphrases should be more or less strict in different scenarios, and really depends on the application to which the paraphrases will be applied.

The task of paraphrasing language which changes over time presents unique challenges as new terms and expressions are invented which are not captured by existing paraphrase systems which capture only a snapshot of common paraphrases at a particular point in time. Although previous work extracting paraphrases from news articles (Dolan et al., 2004) can continuously produce sentences with equivalent meaning from recently written articles, the type of language used in news is quite traditional and thus stable over time. In contrast user-generated text is more likely to reflect new terms and expressions as they are invented (Rumšienė, 2004), so by extracting paraphrases from social media we hope to both automatically and continuously produce up-to-date paraphrase models of the latest terms and expressions as they are invented.

In addition, due to its informal and unconstrained nature, the type of language used in social media presents new challenges and opportunities for paraphrase acquisition. For example, previous corpora of sentential paraphrases (see Section 2 for details) have consisted mostly of declarative sentences. Social media, in contrast, contains paraphrases between questions, declarations, exclamations, and more. As an example, for certain purposes, the following sentences could be considered paraphrases:

- *So will I not be able to fly into NYC Thursday because of Sandy?*
- *Flights into NYC could be canceled on Thursday due to Hurricane Sandy.*

Finally we note that a phrase-based approach is likely required for capturing language

variations across time; for example many phrases in both Shakespearean and Internet English cannot be captured only by word-to-word mappings:

Shakespearean	Standard English	Internet	Standard English
I pray you	please	outta biz	out of business
to tread it	to walk through it	#EGYPT PM	Egyptian Prime Minister

1.2 Paraphrasing Across Language Variations

Little previous work has attempted to model meaning-preserving transformations in different varieties of language. This may be partly because paraphrase data is difficult to obtain, especially between variants of the same language (Xu et al., 2012b). There are only a handful of existing paraphrase corpora, mostly limited to Standard English (Burrows et al., 2012).

1.2.1 Historical Literature and Writing Styles

We investigate the task of automatic paraphrasing while targeting a particular language variety, focusing specifically on the style of Early Modern English employed by William Shakespeare. Besides a relatively good amount of Shakespeare's play scripts, there are many linguistic resources available for us to experiment with various NLP techniques. Among these are parallel "translations" of the plays into colloquial English, as well as dictionaries that provide modern equivalents for archaic words and phrases.

We explore several different paraphrasing methods; some rely on different types of parallel monolingual data techniques from phrase-based MT, and some instead rely on

manually compiled dictionaries of expressions commonly found in Shakespearean English. We evaluate these models both through human judgments and standard evaluation metrics from the Machine Translation (MT) and Paraphrase literature; however no previous work has investigated the ability of automatic evaluation metrics to capture the notion of writing style. We show that previously proposed metrics do not provide a complete picture of a system's performance when the task is to generate paraphrases targeting a specific style of writing. We therefore propose three new metrics for evaluating paraphrases targeting a specific style, and show that these metrics correlate well with human judgments.

1.2.2 Social Media and Internet Language

The emergence of social media (Kwak et al., 2010), e.g. Facebook and Twitter, presents a unique opportunity to collect parallel corpora as it covers a wide range of topics, has a high level of information redundancy, and also contains the most up-to-date language variation and terminology as it is invented (Ke et al., 2008). We extract paraphrases from a parallel corpus consisting of pairs of sentences that are semantically equivalent. In particular, we collect parallel texts from Twitter posts (i.e. tweets) that are related to the same events. Many researchers have attempted to extract paraphrases from news articles about the same events. However, social media has very different characteristics from news articles, motivating us to develop new techniques to cluster events and align words, phrases and sentences.

Meanwhile, the text messages written by millions of users in social networks contain a great deal of valuable and real-time information. It is important to be able to automat-

ically extract and aggregate important information from social networks and Internet language. However, due to the lack of context (e.g. 140-character limit for Twitter) and extremely noisy and flexible nature of user-generated text, many NLP systems suffer from the wide degree of linguistic variations, such as spelling and syntactic variants (or errors), newly-coined terms, etc. Machine learning algorithms trained on static corpora are easily overwhelmed by a large set of novel patterns and terminology. We believe that the ability to identify, extract and generate up-to-date paraphrases would significantly advance the state of the art of natural language processing. We quantitatively investigate the effectiveness of paraphrases in social media data. Our results showed that because tweets include both normal and noisy language, paraphrase systems built from Twitter could be fruitfully applied to the task of normalizing noisy text, covering phrase-based normalizations not handled by state-of-the-art dictionary-based normalization systems.

We expect that paraphrases would help to transfer existing knowledge, annotations and NLP tools for Standard English to noisy user-generated text. Focusing on the emergent social media, which becomes an important part of people's everyday lives and practices, our research will help scientists and humanists from sociology, communications, anthropology, media studies, information science, medical science and cultural studies (Chunara et al., 2012; Eisenstein et al., 2010; Naaman et al., 2011; Xu et al., 2012a).

In conclusion, modeling paraphrases for language variations will extend automatic language processing to languages of different time periods, and assist people in understanding and writing such languages.

1.3 Overview

The rest of this thesis is organized as follows. We begin with a review of prior work in the related areas in Chapter 2.

To demonstrate the feasibility and value of paraphrasing language variations, we first approach the task of transforming the plays of William Shakespeare into modern English and vise versa, and present several different paraphrasing models that utilize various language resources in Chapter 3. During the evaluation of these models, we find that none of the previous evaluation metrics from the Machine Translation (MT) and Paraphrase literature can provide the complete picture of a system's performance when the task is to generate paraphrases targeting a specific style of language. We therefore propose three new metrics for evaluating paraphrases targeting a specific writing style, and show that these metrics correlate well with human judgments in Chapter 4.

We next turn to Internet language, in particular Twitter, which contains the most up-to-date information and linguistic phenomena. In Chapter 5, we show how to paraphrase informal user-generated text using a parallel corpus automatically gathered from Twitter by information extraction techniques. We further demonstrate the effectiveness of these paraphrase models to normalize noisy text. Although the corpus collected in Chapter 5 is at very large scale, it is limited by the constraints and errors associated with the automatic processes. We leverage crowd-sourcing to gather cleaner data (Chapter 6) and train systems for paraphrase identification on Twitter (Chapter 7).

Finally in Chapter 8, we conclude the thesis and discuss future works.

Throughout the thesis, we use the term *style* to simply refer to the characteristics of a variation of language. As in the literature of natural language processing (Jiang, 2008;

Li, 2012), the notion of *domain* is typically inclusive of significant language variations, genres or topics etc. Social media can be considered as a single domain that contains both formal and informal languages; while the news from the same news sources that cover Middle East and United States can be considered as different domains. We use the term *noisy* in two different meanings in this thesis. One is to stress the characteristic of social media text (Baldwin et al., 2013), that includes misspellings, grammar errors, word-choice errors, creative abbreviations and etc. The other one means the errors in the annotated data (Spreyer and Kuhn, 2009), such as a non-paraphrase sentence pair mislabeled as paraphrase in the corpus.

Chapter 2

Related Work

The paraphrase problem is fundamental to many natural language understanding tasks (Giampiccolo et al., 2007), therefore much work has investigated methods for automatic paraphrasing. We are not aware, however, of any work that has systematically addressed the task of extracting, generating or evaluating paraphrases targeting variations of language. Most of the existing work deals with texts from the same domain or from different domains which do not contain as great linguistic differences as ours. To some extent, our task is close to machine translation, from which we also draw inspiration. This chapter highlights the contribution of this thesis in the context of the most related previous work. For a more comprehensive background on paraphrasing, see the excellent surveys by Madnani and Dorr (2010), Burrows et al. (2012) and Ho et al. (2012).

2.1 Paraphrase Acquisition

Automatic Paraphrase Acquisition approaches have emerged and become extremely popular in the last decade (Bannard and Callison-Burch, 2005; Barzilay and Lee, 2003; Das and Smith, 2009; Dolan et al., 2004; Shinyama and Sekine, 2003). However, as noted by (Dolan and Brockett, 2005), a paraphrase corpus is difficult to obtain, "since paraphrase is not apparently a common natural task – under normal circumstances people do not attempt to create extended paraphrase texts".

One data source has been different translations or versions of the same text (Barzilay and McKeown, 2001; Ibrahim, 2002; Ibrahim et al., 2003). Some researchers exploit a foreign language as a pivot to extract paraphrase patterns from bilingual parallel corpora (Callison-Burch, 2008; Ganitkevitch et al., 2013; Zhao et al., 2008). However, bilingual parallel resources are not readily available for social media. Only very recently, a study which makes use of bilingual text from Chinese microblog, Sina Weibo, has been reported by Wang et al. (2013).

Previous work has also investigated the tasks of gathering parallel or comparable texts from monolingual resources (Barzilay and McKeown, 2001; Clough et al., 2002; Cohn et al., 2008; Dolan et al., 2004; Dolan and Brockett, 2005; Fader et al., 2013; Knight and Marcu, 2002; Lin and Pantel, 2001; Paşca and Dienes, 2005; Sekine, 2005; Shinyama and Sekine, 2003; Shinyama et al., 2002). Very related to our approach for collecting paraphrases is previous work by Shinyama and Sekine (2003), Dolan et al. (2004) and Quirk et al. (2004). This line of work aims to extract paraphrases from news articles written by different press agencies describing the same event. The intuition behind our approach is very similar: many social media users will refer to the same events. Social

media has very different characteristics from news articles, however, motivating us to develop different techniques for clustering events and sentence alignment.

There are only a few recent studies on paraphrasing between slightly different domains, such as technical and lay medical terms (Deléger and Zweigenbaum, 2009; Elhadad and Sutaria, 2007). The variations between languages we are dealing with are much greater than theirs.

Crowd Paraphrase Acquisition has been investigated recently, such as in (Buzek et al., 2010) and (Denkowski et al., 2010). Both focus specifically on collecting paraphrases of text to be translated to improve the quality of machine translations from crowdsourcing services. Chen and Dolan (2011) gathered a large-scale sentence-level paraphrase corpus by asking independent users of Amazon's Mechanical Turk Service (Snow et al., 2008) to caption the action in short video segments. Similarly, Burrows et al. (2012) collected passage-level paraphrases by asking crowdsourcing workers to rewrite excerpts chosen from books.

By contrast, we design a simple task on a crowdsourcing platform requesting only binary judgement on sentences that are topically and temporally related from Twitter (Chapter 6). There are several important differences and advantages as compared to other existing work: 1) this method is scalable and sustainable due to the simplicity of the task and real-time unlimited text supply from Twitter; 2) the paraphrase corpus collected contains a representative proportion of both negative and positive instances, while lack of good negative examples was an issue in the previous research (Das and Smith, 2009); 3) the corpus also covers a very diverse range of topics and linguistic expressions, especially colloquial language, which is different from and thus complements previous

paraphrase corpora. There exist two sentence-level paraphrase corpora that are of an order of magnitude that is similar to ours (more than 5000 sentence pairs). Table 2.1 illustrates the different characteristics of the data we collected and those produced by previous work.

Corpus	**Examples**
Microsoft (Dolan and Brockett, 2005)	○ Revenue in the first quarter of the year dropped 15 percent from the same period a year earlier. ○ With the scandal hanging over Stewart's company, revenue in the first quarter of the year dropped 15 percent from the same period a year earlier.
	○ The Senate Select Committee on Intelligence is preparing a blistering report on prewar intelligence on Iraq. ○ American intelligence leading up to the war on Iraq will be criticized by a powerful US Congressional committee due to report soon, officials said today.
Video (Chen and Dolan, 2011)	○ A person is slicing a cucumber into pieces. ○ A man cutting zucchini.
	○ Someone is coating a pork chop in a glass bowl of flour. ○ A person breads a pork chop.
Twitter (This Work)	○ the utmost respect for Harding ○ So impressed by Harding s effort
	○ Ezekiel Ansah wearing 3D glasses wout the lens. ○ Ezekiel Ansah 5th overall pick wearing real3D glasses with the lenses popped out.

Table 2.1: Representative examples from paraphrase corpora

2.2 NLP for Social Media

Social media websites provide a massive amount of timely and important information, motivating the need for language technology. Standard text processing tools perform poorly when applied to social media text due to its noisy and unique language (Gimpel et al., 2011; Ritter et al., 2011b), which is very different from traditional sources (e.g. newswire). Several researchers have attempted to address the problem by designing new algorithms specifically for social media data, including information extraction (Ritter et al., 2011b, 2012), retrieval (Subramaniam et al., 2009), summarization (Chakrabarti and Punera, 2011; Liu et al., 2011b), sentiment analysis (Celikyilmaz et al., 2010), semantic role labeling (Liu et al., 2010, 2011c) and first story detection (Petrović et al., 2012). Yet their performance is still limited by the lack of domain knowledge and annotated data. Paraphrasing from social media texts to their most likely Standard English versions will help to leverage existing knowledge and techniques. Parallel aligned text also opens the possibility for transferring annotations as in bilingual research (Yarowsky et al., 2001).

Our research is also related to previous work on normalizing Twitter text (Han and Baldwin, 2011; Han et al., 2012; Liu et al., 2011a; Liu et al., 2012), which is limited to spelling corrections of non-standard English tokens. It extends the idea of exploiting large-scale web data for word choice in our previous work (Xu et al., 2011a). Access to parallel text allows us to achieve a much broader objective, which is to acquire phrasal paraphrases and generate sentential paraphrases. Very related to our work is the recent paper by Wang et al. (2013), which treats text normalization as a paraphrasing task as we do but is limited by the availability of bilingual texts in social media. Zanzotto et al. (2011) also reported an initial study on detecting redundant posts in Twitter.

2.3 Statistical Machine Translation

Many statistical machine translation (SMT) techniques have been successfully applied to paraphrasing problems, which can be seen as monolingual translations, through parallel corpora. One of the crucial keys to this approach is to align all the corresponding text fragments together: align sentence pairs to create a parallel corpus and align word or phrase pairs to extract paraphrases. To extract parallel sentences from a monolingual comparable corpus, most previous work relies on lexical similarity, which is ineffective for the very disparate data that is the subject of our study. Our approach is inspired by previous work on bilingual comparable corpora (Fung and Cheung, 2004a,b; Lee et al., 2010; Smith et al., 2010), while answering the new challenges in the noisy social media data.

2.4 Domain Adaptation

The focus of our study, language variations over time, is outside the typical scope of domain adaption in natural language processing. Despite many studies addressing domain adaptation from different perspectives (Bacchiani and Roark, 2003; Blitzer et al., 2006; Daumé III and Marcu, 2006; Jiang and Zhai, 2006; Mansour et al., 2009), most of them focus on differences either between different datasets (i.e. both newswire), genres (i.e. newswire and terrorists reports) or subsets within the same dataset. The linguistic differences between these domains are certainly not as extensive and structural as those developed in language evolution, which we are facing. Most previous work adapted NLP tools to Twitter by using annotated in-domain data or specially designed features

(Gimpel et al., 2011; Ritter et al., 2011b). Most relevant to our study is the previous work by Liu et al. (2010) leveraging a semantic role labeling system on the news domain to tweets of news content. By modeling paraphrases, we can provide a more generic framework of domain adaptation that is not limited to a certain application or a certain genre. Paraphrasing also goes beyond domain adaptation in that it could help people with writing and understanding different varieties of language.

Chapter 3

Shakespearean Paraphrasing

Motivated by the lack of computational study on language changes, we conducted initial investigation into the task of paraphrasing between varieties of language using Shakespeare's plays as a testbed (Xu et al., 2012b). Because these plays are some of the most highly-regarded examples of English literature and are written in a style that is now 400 years out of date, many linguistic resources are available to help modern readers better understand these Elizabethan texts. We show that even with a relatively small amount of parallel training data, it is possible to learn paraphrase models which capture linguistic phenomena by an approach based on phrase-based statistical machine translation, and these models outperform baselines based on dictionaries and out-of-domain[1] paraphrase corpora. This work validated our hypotheses that it is feasible to learn paraphrases between variations of one language from a parallel corpus, and it is more efficient than using lexical recourses.

1. In this chapter, the Shakespeare's plays and their modern translations are considered in-domain; while other texts, e.g. news and video descriptions, are out-of-domain.

3.1 Paraphrasing into Shakespearean English

We compared three different paraphrase systems targeting Shakespearean English which rely on different types of linguistic resources. One leverages parallel translations, another exploits dictionary resources, and a third relies on modern, out-of-domain monolingual parallel data and an in-domain language model.

3.1.1 Parallel Modern Translations

Access to parallel text in the target style allows us to train statistical models that generate paraphrases, and also perform automatic evaluation of semantic adequacy using BLEU, which requires availability of reference translations. For this purpose we scraped modern translations of 17 Shakespeare's plays from http://nfs.sparknotes.com, and additional translations of 8 of these plays from http://enotes.com.

After tokenizing and lowercasing, the plays were sentence aligned (Moore, 2002), producing 21,079 alignments from the 31,718 sentence pairs in the Sparknotes data, and 10,365 sentence pairs from the 13,640 original pairs in the Enotes data. The modern translations from the two sources are qualitatively quite different. The Sparknotes paraphrases tend to differ significantly from the original text, whereas the Enotes translations are much more conservative, making fewer changes. To illustrate these differences empirically and provide an initial paraphrase baseline, we compute BLEU scores of the modern translations against Shakespeare's original text; the Sparknotes paraphrases yield a BLEU score of 24.67, whereas the Enotes paraphrases produce a much higher BLEU of 52.30 reflecting their strong similarity to the original texts. These results are summarized in Table 3.1.

corpus	initial size	aligned size	No-Change BLEU
http://nfs.sparknotes.com	31,718	21,079	24.67
http://enotes.com	13,640	10,365	52.30

Table 3.1: Parallel corpora generated from modern translations of Shakespeare's plays

Phrase-based translation has been demonstrated as an effective approach to generate paraphrases (Chen and Dolan, 2011; Quirk et al., 2004). We applied a typical phrase-based statistical MT pipeline, performing word alignment on the data described in table 3.1 using GIZA++ (Och and Ney, 2003), then extracting phrase pairs and performing decoding using Moses (Koehn et al., 2007). We used this pipeline with the basic default settings recommended in the toolkits documentation[2,3]. All the language models used in this thesis are trained on different datasets by SRILM toolkit (Stolcke, 2002) with trigrams and interpolated interpolated Kneser-Ney smoothing[4], if not otherwise specified.

For evaluation purposes, the parallel text of one play, Romeo and Juliet, was held out of the training corpus for this system and the baseline systems described in the following section.

3.1.2 Dictionary Based Paraphrase

The statistical machine translation approach does require the existence of parallel corpora of aligned phrases and sentences, however, resources which may not be available for many language variations that we might wish to target. For this reason we were

2. http://www.statmt.org/moses_steps.html
3. http://www.statmt.org/moses/?n=Moses.Baseline
4. The SRILM command line parameters we used are "ngram-count -order 3 -interpolate -kndiscount -text inputtexfilename -lm outputlmfilename"

motivated to investigate alternative approaches in order to help quantify how critical this type of parallel data is for the task of stylistic paraphrasing.

Several dictionaries of stylistically representative words of Shakespearean English and their modern equivalents are available on the web. These dictionaries can be used to define a translation model which can be used in combination with a language model as in standard phrase-based MT.

To build a phrase table, we scraped a set of 68,709 phrase/word pairs from http://www.shakespeareswords.com/; example dictionary entries are presented in table 3.2. As described in (Koehn and Knight, 2000), we estimate phrase translation probabilities based on the frequencies of the translation words/phrases in the target language (Shakespearean English). For instance, if we look at the modern English word *maybe*, our dictionary lists 4 possible Shakespearean translations. We obtained the probabilities for each translation according to the n-gram back-off model built from 36 of Shakespeare's plays using the SRILM toolkit (Stolcke, 2002), normalizing the probabilities for each source phrase, for example $p(\text{PERCHANCE}|\text{maybe}) = \frac{0.000079}{0.000263} = 0.301$. An example is presented in Table 3.3. This method allows us to estimate reasonable translation probabilities for use in a phrase table, which is used in combination with a language model built from the 36 plays, which are then fed into the Moses decoder (Koehn et al., 2007). For out-of-the-vocabulary words in the input, we set the Moses to pass them through.

target	source	target	source
ABATE	shorten	AYE	always
CAUTEL	deceit	GLASS	mirror
SUP	have supper	VOICE	vote

Table 3.2: Example dictionary entries

Smoothed Probability Estimate	target	source
0.0000791	PERCHANCE	maybe
0.0000369	PERADVENTURE	maybe
0.0000752	HAPLY	maybe
0.0000714	HAPPILY	maybe
total 0.000263		

Table 3.3: Example ngram probabilities in target language

3.1.3 Out of Domain Monolingual Parallel Data

As a final baseline we consider a paraphrase system which is trained on out-of-domain data gathered by asking users of Amazon's Mechanical Turk Service (Snow et al., 2008) to caption the action in short video segments (Chen and Dolan, 2011). We combined a phrase table extracted from this modern, out of domain parallel text, with an in-domain language model consisting of Shakespeare's 36 plays, applying the Moses decoder (Koehn et al., 2007) to find the best paraphrases. Although this monolingual parallel data does not include text in the target writing style, the in-domain language model does bias the system's output towards Shakespeare's style of writing. We found that performing Minimum Error Rate Training (Och, 2003) using a small set of held out parallel text from Romeo and Juliet was necessary in order to tune the video corpus baseline to generate reasonable paraphrases.

3.2 Automatic Evaluation

Figure 3.1 compares a variety of systems targeting Shakespearean English using the previously proposed BLEU (Papineni et al., 2002) and PINC (Chen and Dolan, 2011) automatic evaluation metrics which have been demonstrated to correlate with human judgments on semantic adequacy and lexical dissimilarity with the input. A description of each of the systems compared in this experiment is presented in Table 3.4. As mentioned in §3.1.1, the Enotes paraphrases diverge little from the original text, resulting in a BLEU score of 52.3 when compared directly to the original lines from Shakespeare's plays. Because our goal is to produce paraphrases which make more dramatic stylistic changes to the input, in the remainder of this paper, we focus on the Sparknotes data for evaluation.

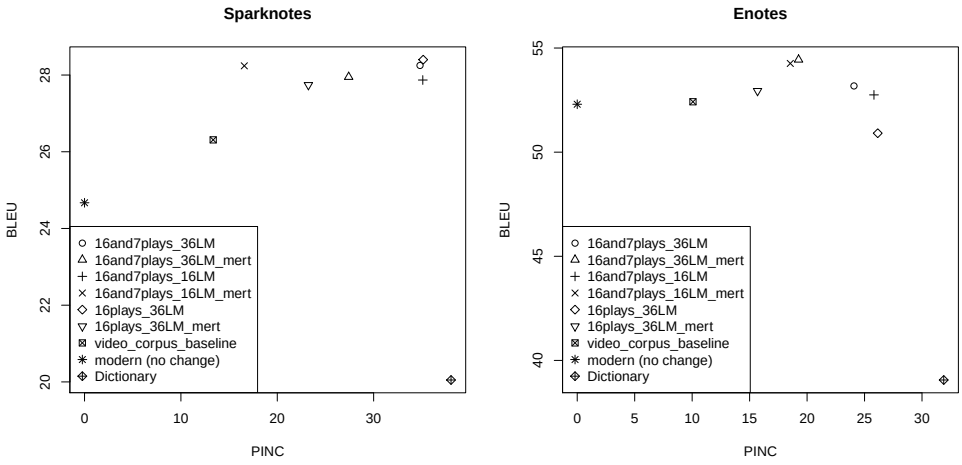

Figure 3.1: Various Shakespearean paraphrase systems compared using BLEU and PINC. A brief description of each system is presented in table 3.4.

Two main trends are evident in Figure 3.1. First, notice that all of the systems trained

System	Description
16and7plays_36LM	Phrase table extracted from all 16 Sparknotes plays and 7 Enotes plays (holding out R&J) and language model built from all 36 of Shakespeare's plays, again excluding R&J. Uses default Moses parameters.
16and7plays_36LM_MERT	Same as 16and7plays_36LM except parameters are tuned using Minimum Error Rate Training (Och, 2003).
16and7plays_16LM	Phrase table is built from both Sparknotes and Enotes data, and Language model is built from the 16 plays with modern translations
16and7plays_16LM_MERT	Same as 16and7plays_16LM except parameters are tuned using MERT.
16plays_36LM	Only Sparknotes modern translations are used. All 36 plays are used to train Shakespearean language model.
16plays_36LM_MERT	Same as 16plays_36LM except parameters are tuned using MERT.
video_corpus_baseline	Paraphrase system combining out of domain parallel text (Chen and Dolan, 2011) with an in-domain language model. Described in detail in §3.1.3.
modern (no change)	No changes are made to the input, modern translations are left unchanged.
Dictionary	Dictionary baseline described in §3.1.2

Table 3.4: Descriptions of various systems for Shakespearean paraphrase. *Romeo and Juliet* is held out for testing.

using parallel text achieve higher BLEU scores than the unmodified modern translations. While the dictionary baseline achieves a competitive PINC score, indicating it is making a significant number of changes to the input, its BLEU is lower than that of the modern translations. Secondly, it seems apparent that the systems whose parameters are tuned

using Minimum Error Rate Training (MERT) tend to be more conservative, making fewer changes to the input and thus achieving lower PINC scores, while not improving BLEU on the test data. Finally we note that using the larger target language model seems to yield a slight improvement in BLEU score.

Example paraphrases of lines from *Romeo and Juliet* and several Hollywood movies, generated by the top performing system according to BLEU and PINC, are presented in table 3.5.

3.3 Human Evaluation

Figure 3.1 provides some insight into the performance of the various systems, but it is initially unclear how well the BLEU and PINC automatic evaluation metrics perform when applied to paraphrases that target a specific style of writing. BLEU and PINC have previously been shown to have high correlation with human judgments of semantic adequacy and lexical dissimilarity of paraphrase candidates, but the implications of this for the more specialized task of stylistic paraphrasing are unclear.

While BLEU is typically used to measure semantic adequacy, it seems reasonable to assume that it could also be useful for measuring stylistic alternations, since utterances are more likely to contain overlapping ngrams if they are both semantically and stylistically similar. What BLEU cannot tell us, however is what portion of its improvements are due to stylistic similarity or semantic equivalence. For this reason, we were motivated to perform an evaluation based on human judgments of semantic adequacy, lexical dissimilarity and stylistic similarity.

For this purpose, we randomly sampled 100 lines from *Romeo and Juliet*, then two

Source	Speaker	Input	Output
Romeo & Juliet	Benvolio	He killed your relative, brave Mercutio, and then young Romeo killed him.	he slew thy kinsman , brave mercutio , and then young romeo kill him .
Romeo & Juliet	Romeo	I can read my own fortune in my misery.	i can read mine own fortune in my woes .
Star Wars	Palpatine	If you will not be turned, you will be destroyed!	if you will not be turn 'd , you will be undone !
Star Wars	Luke	Father, please! Help me!	father , i pray you , help me !
The Matrix	Agent Smith	Good bye, Mr. Anderson.	fare you well , good master anderson .
The Matrix	Morpheus	I'm trying to free your mind, Neo. But I can only show you the door. You're the one that has to walk through it.	i 'll to free your mind , neo. but i can but show you the door. you 're the one that hath to tread it .
Raiders of the Lost Ark	Belloq	Good afternoon, Dr. Jones.	well met , dr. jones .
Raiders of the Lost Ark	Jones	I ought to kill you right now.	i should kill thee straight .
Terminator 2	Terminator	My mission is to protect you.	my affair is to keep you .
Terminator 2	Terminator	Negative. The T-1000's highest probability for success now would be to copy Sarah Connor and to wait for you to make contact with her.	negative. the t-1000 's most exalted probability for success now would be to have sarah connor and to stay for you to make contact with her .
Terminator 2	Terminator	Skynet begins to learn at a geometric rate. ... In a panic, they try to pull the plug.	skynet begins to learn at a geometric rate. ... in a fearful sails , they would pluck the stop .

Table 3.5: Example Shakespearean paraphrases generated by the best overall system.

annotators annotated each sentence and its Shakespearean translation to indicate semantic adequacy, lexical dissimilarity, stylistic similarity, and overall quality. The aggregate results of the human evaluation are displayed in Figure 3.2. Agreement between annotators measured using Pearson's ρ is displayed in Table 3.6.

Based on the human evaluation, it appears that the baseline combining paraphrases collected from Mechanical Turk (Chen and Dolan, 2011) with a Shakespearean language model has the highest semantic adequacy, yet this approach is also fairly conservative in that it makes few changes to the input.

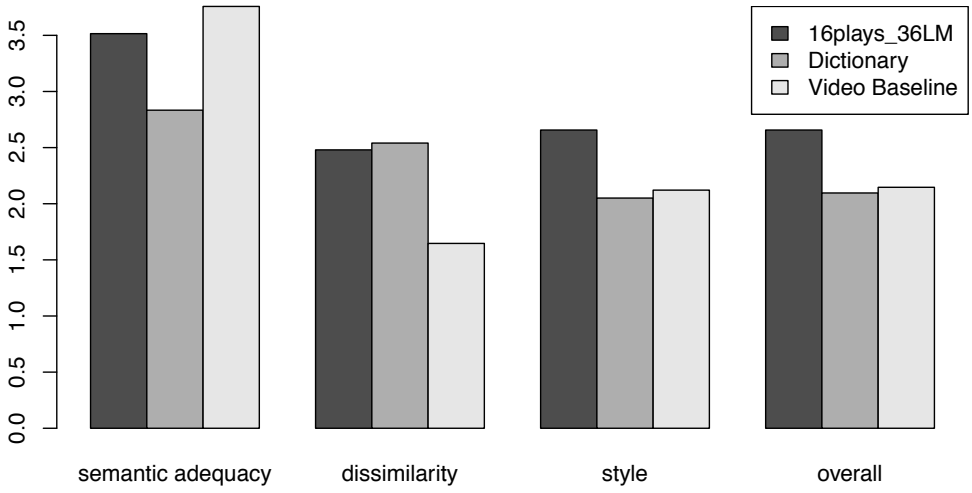

Figure 3.2: Average human judgments evaluating semantic adequacy, lexical dissimilarity, stylistic similarity, and overall quality of Shakespearean paraphrase systems

Semantic Adequacy	Lexical Dissimilarity	Style	Overall
0.73	0.82	0.64	0.62

Table 3.6: Agreement between annotators measured using Pearson's ρ.

The dictionary baseline, and the paraphrase system trained on parallel modern translations are roughly comparable in terms of the number of changes made to the input, but the system trained on modern translations achieves higher semantic adequacy, while also being rated higher on style and overall.

These results are roughly in line with the automatic metrics presented in Figure 3.1. However we also see several important trends which are not apparent from the automatic evaluation. Although the video baseline achieves the highest semantic adequacy in the human evaluation, its BLEU score is significantly lower than 16plays_36LM on the Sparknotes data.[5] It would appear that in this case BLEU is conflating semantic adequacy with writing style. Although the paraphrases produced by the video baseline have high semantic adequacy, their style tends to differ substantially from the reference translations resulting in fewer ngram matches, and thus a lower BLEU score.

While existing evaluation metrics do seem useful for evaluating stylistic paraphrases, they are not capable to separate the notion of writing style from semantic adequacy. This motivated us to develop automatic evaluation metrics to distinguish between a system which generates perfect paraphrases which do not match the target style of writing versus a system which generates sentences in the correct style, but which convey different meaning, which is subsequently presented in the Chapter 4.

5. Note that the BLEU score of 16plays_36LM is significantly lower when evaluated on the Enotes data. This makes sense, because the 16 plays come from Sparknotes. This system is not trained on the 7 Enotes plays, whose modern translations tend to be slightly different in style.

3.4 Translating Shakespeare's Plays to Modern English

We also perform an evaluation on the task of automatically translating Shakespeare's plays into modern English. We make use of the same paraphrase systems previously described, but swap the source and target languages. Additionally, each system makes use of a language model constructed from the 16 modern translations, with *Romeo and Juliet* held out for testing. 100 lines from *Romeo and Juliet* were automatically translated into modern English using each system, and the aligned modern translations were used as a reference when computing BLEU. The results of evaluating each of the automatic evaluation metrics on this data are presented in Figure 3.3 and average human judgments are presented in Figure 3.4.

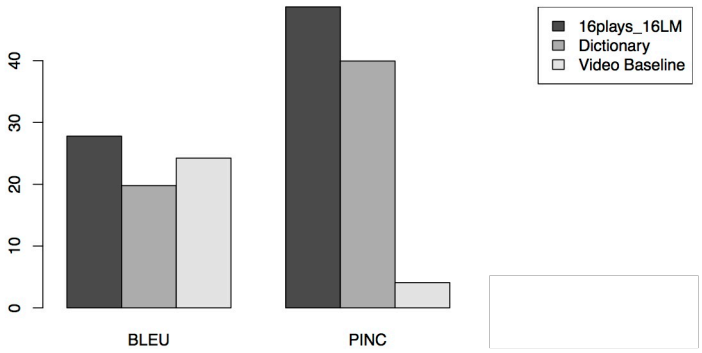

Figure 3.3: Automatic evaluation of paraphrasing Shakespeare's plays into modern English comparing a system based on parallel text (16plays_16LM), a Dictionary baseline, and a system trained on out of domain parallel monolingual text.

These results suggest that in comparison to the dictionary and video corpus baselines,

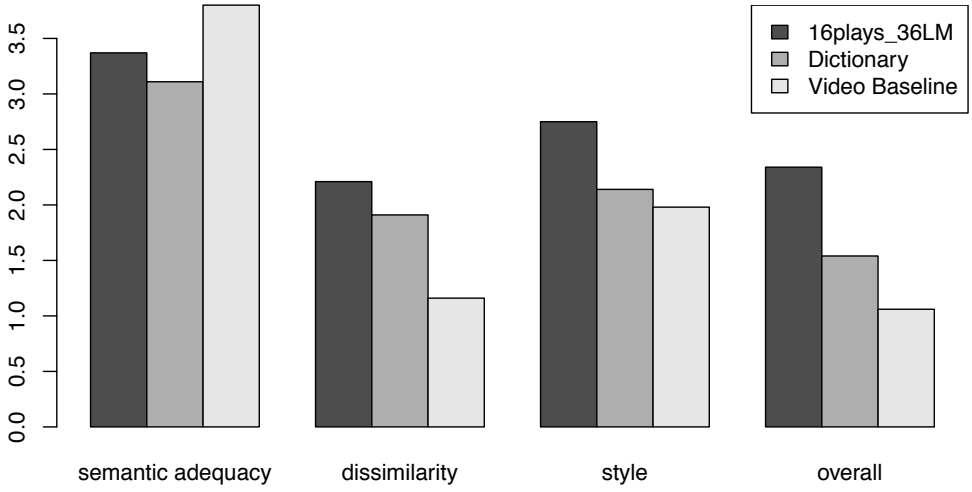

Figure 3.4: Average human judgments translating Shakespeare's plays into modern English.

our system trained on modern translations generates a large number of paraphrases which match the target style. Note that the paraphrase system based on the out-of-domain video corpus makes very few changes to the input, and thus achieves a very low PINC score. This is due to the many out of vocabulary words in Shakespeare's plays which result in very few matching source phrases in the video baseline's phrase table. Several automatic paraphrases into modern English are presented in Table 3.7.

3.5 Related Work

Much previous work has addressed the task of automatically generating paraphrases (Bannard and Callison-Burch, 2005; Barzilay and Lee, 2003; Callison-Burch, 2008; Das and Smith, 2009; Dolan et al., 2004; Kok and Brockett, 2010; Shinyama and Sekine, 2003). There is also a significant literature in sentence compression aimed at modeling

Speaker	Input	Output
MERCUTIO	i will bite thee by the ear for that jest .	i ' ll bite you by the ear for that joke .
MONTAGUE	what further woe conspires against mine age ?	what ' s true despair conspires against my old age ?
ROMEO	how doth my lady ?	how is my lady ?
FRIAR LAURENCE	hast thou slain tybalt ?	have you killed tybalt ?
NURSE	an i might live to see thee married once , i have my wish .	if i could live to see you married , i ' ve my wish .
PRINCE	benvolio , who began this bloody fray ?	benvolio , who started this bloody fight itself ?
JULIET	what is your will ?	what do you want ?
LADY CAPULET	call her forth to me .	bring her out to me .

Table 3.7: Example modern paraphrases of lines from Romeo and Juliet generated using our system.

the first of these, length: producing meaning-preserving alternations that reduce the length of the input string (Chandrasekar et al., 1996; Clarke and Lapata, 2008; Cohn and Lapata, 2009; Ganitkevitch et al., 2011; Vanderwende et al., 2007; Xu and Grishman, 2009; Yatskar et al., 2010). We are not aware, however, of any work that has addressed the task of generating or evaluating paraphrases targeting a specific style of writing.

Perhaps most relevant, however, is recent work on automatic generation of rhythmic poetry (Greene et al., 2010). This work focuses on automatically generating and translating poetry in an appropriate meter (e.g. iambic pentameter) using finite-state transducers, but does not investigate the task of paraphrase. Their generation system is trained on Shakespeare's sonnets, and they investigate the task of automatically translating Dante's Divine Comedy from Italian to English. While our work does not address

the issue of meter, it should be possible to combine our translation models with their weighted finite state transducers to produce Shakespearean paraphrase models which produce output in an appropriate meter.

We also note a very recent study that learns a machine translation system that accepts hip hop lyric challenges and improvises rhyming responses (Wu et al., 2013).

Finally we highlight related work on authorship classification which can be seen as detecting a specific style of writing (Gamon, 2004; Raghavan et al., 2010). This work has not specifically addressed the task of automatically generating or evaluating paraphrases in a specific style, however.

3.6 Conclusions

We have presented the first investigation into the task of automatic paraphrasing while targeting a specific writing style. Using Shakespeare's plays and their modern translations as a testbed for this task, we developed a series of paraphrase systems targeting Shakespearean English. We have shown that access to even a small amount of parallel text produces paraphrase systems capable of generating a large number of stylistically appropriate paraphrases while preserving the meaning of the input text. Our paraphrase systems targeting Shakespearean English could be beneficial for educational applications, for example helping to make Shakespeare's work accessible to a broader audience.

We also demonstrated that existing evaluation metrics developed in the Machine Translation and Paraphrase communities are insufficient when the goal is to generate paraphrases targeting a specific style. So in the next chapter, we propose a series of new

metrics to measure how closely the generated paraphrases match the target writing style.

Chapter 4

Automatic Metrics Evaluating Writing Style

As we showed in the last chapter, while existing evaluation metrics are useful for evaluating paraphrases, they cannot differentiate stylistic similarity from semantic equivalence and thus give an incomplete picture of system performance when the task is to generate paraphrases targeting a specific style. To help address this issue we propose three new automatic evaluation metrics whose goal is to measure the degree to which automatic paraphrases match the target style. These metrics assume existence of large corpora in both the source and target style, but do not require access to any parallel text, or human judgments. We present a preliminary evaluation of the proposed metrics by measuring their correlation with human judgments, but it should be emphasized that we are only evaluating these metrics with respect to two specific styles of writing, Shakespearean vs. contemporary Standard English. We are optimistic that these results will generalize across writing styles, however, since they are based entirely on ngram

statistics.

4.1 Cosine Similarity Style Metric

As a first approach to automatic evaluation of writing style, we present a vector-space model of similarity between the system output and a large corpus of text in both the source and target style. The intuition behind this metric is that a large ngram overlap between the system's output and a corpus of text in the target style should indicate that the output is likely to be stylistically appropriate.

More concretely, we extract ngrams from both the source and target corpus which are represented as binary vectors $\vec{s}$, and $\vec{t}$; similarly the output sentence is represented using a vector of ngrams $\vec{o}$. The proposed metric is the normalized cosine similarity between the source and target corpora:

$$S_{\text{Cosine}}(\vec{o}) = \frac{\frac{\vec{o}\cdot\vec{t}}{\|\vec{o}\|\times\|\vec{t}\|}}{\frac{\vec{o}\cdot\vec{t}}{\|\vec{o}\|\times\|\vec{t}\|} + \frac{\vec{o}\cdot\vec{s}}{\|\vec{o}\|\times\|\vec{s}\|}}$$

4.2 Language Model Style Metric

Another approach is to build a language model from a corpus of text in the target style and a background language model from text outside the style, then apply Bayes' rule to estimate the posterior probability that a sentence was generated from the target

language model[1]:

$$\begin{aligned} P(\text{style} = \text{target}|\text{sentence}) &= \frac{P_{\text{LM}}(\text{sentence}|\text{target})P(\text{target})}{P(\text{sentence})} \\ &= \frac{P_{\text{LM}}(\text{sentence}|\text{target}) \times 0.5}{P_{\text{LM}}(\text{sentence}|\text{target}) \times 0.5 + P_{\text{LM}}(\text{sentence}|\text{source}) \times 0.5} \\ &= \frac{P_{\text{LM}}(\text{sentence}|\text{target})}{P_{\text{LM}}(\text{sentence}|\text{target}) + P_{\text{LM}}(\text{sentence}|\text{source})} \end{aligned}$$

4.3 Logistic Regression Style Metric

We also consider an approach to measuring style which is based on logistic regression. Here the idea is to estimate the probability that each sentence belongs to the target style based on the ngrams it contains, using two large corpora in the target and source styles to learn parameters of a logistic regression model.

The probability that a sentence belongs to the target style is estimated as follows:

$$P(\text{style} = \text{target}|\text{sentence}) = \frac{1}{1 + e^{-\left(\vec{\theta} \cdot f(\vec{\text{sentence}})\right)}}$$

Where $f(\vec{\text{sentence}})$ is a vector of ngrams contained by the sentence, and $\vec{\theta}$ is a vector of weights corresponding to each possible ngram.

The parameters, $\vec{\theta}$, are optimized to maximize conditional likelihood on the source and target corpus, where the assumption is that the target corpus is in the target style, whereas the source corpus is not.[2]

1. Here we assume an uninformative prior, that is $P(\text{source}) = P(\text{target}) = 0.5$.
2. Parameters were optimized using MEGAM http://www.cs.utah.edu/-hal/megam/.

4.4 Evaluation

We conduct our experiments on the same data described in §3.1.1, which includes 36 Shakespeare's plays and modern translations of 17 plays from http://nfs.sparknotes.com. For the language model style metric, we use the interpolated back-off trigram language model, which is also described in details in §3.1.1,

4.4.1 Measuring Shakespearean Style

We first evaluate on the task of paraphrasing while targeting the Shakespearean style. We trained the logistic regression, language model and cosine similarity evaluation metrics using the original Shakespeare plays and modern translations as the source and target corpus respectively, then measured Pearson's Correlation Coefficient between the automatic evaluation metrics and human judgments on the 100 seperate lines of *Romeo and Juliet* described in §3.3. These results are reported in table 4.1.

		ρ (Annotator 1)	ρ (Annotator 2)
semantic adequacy	BLEU	0.35	0.31
dissimilarity	PINC	0.78	0.82
style	BLEU	0.07	0.06
style	PINC	0.20	0.45
style	Cosine	0.37	0.41
style	LM	0.46	0.51
style	Logistic regression	0.47	0.47

Table 4.1: Correlation between various human judgments and automatic evaluation metrics. Pearson's correlation coefficient is displayed between the automatic metrics and human judgments from each annotator.

As can be seen in table 4.1, the correlation between semantic adequacy and BLEU

appears smaller than that reported in previous work (Chen and Dolan, 2011). Presumably this is due to the conflation of stylistic differences and semantic adequacy discussed in §3.3. However it also appears that the correlation between BLEU and human style judgments is too low to be of practical use for evaluating style.

PINC, on the other hand has high correlation with judgments on dissimilarity, and is also correlated with human style judgments. We believe PINC has correlation with writing style, because the systems we are evaluating all target Shakespearean English, so whenever changes are made to the input, they are likely to make it similar to the target style. Although PINC has relatively high correlation with human judgments, it is likely not a very useful measure of writing style in practice. For example, consider a paraphrase system which makes many changes to the input and thus gets a high PINC score, but targets a completely different writing style.

Both the language model and logistic regression style metrics achieve the highest overall correlation with human writing style judgments, achieving comparable performance.

We note that overall the automatic metrics (Figure 4.1) tend to agree with human judgments (Figure 3.2).[3]

4.4.2 Measuring Modern Prose

Similarly, we perform an evaluation on the opposite direction on the same 100 lines from *Romeo and Juliet*, automatically transforming Shakespeare's plays into modern English.

3. Although the automatic style metrics rate the dictionary system higher than the video corpus baseline, both systems have very comparable style scores in the automatic and human evaluations.

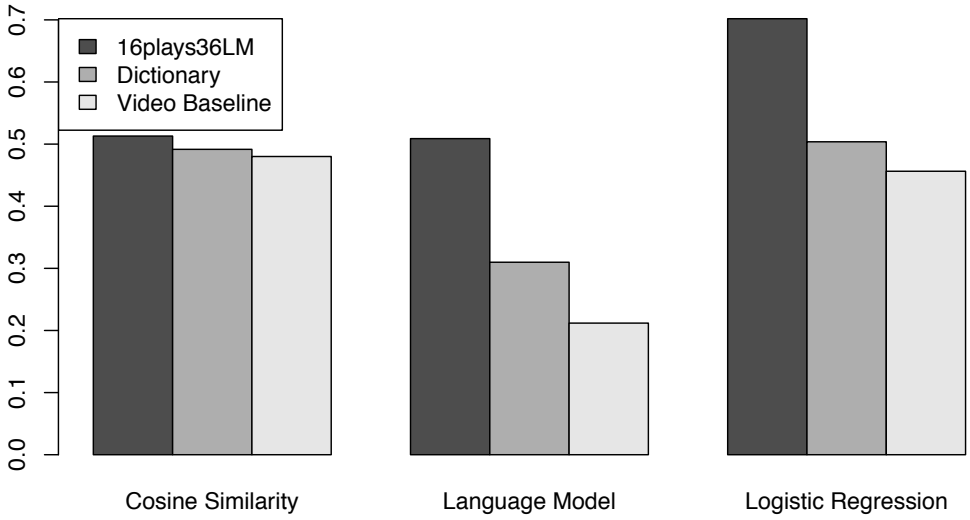

Figure 4.1: Automatic stylistic evaluation comparing the three systems that paraphrase into Shakespearean style.

The results of evaluating each of the automatic evaluation metrics on this data are presented in Figure 4.2 and tend to agree with human judgements as displayed in Figure 3.4. The correlation of the automatic metrics with human judgments are presented in Table 4.2.

		ρ (Annotator 1)
semantic adequacy	BLEU	0.27
dissimilarity	PINC	0.79
style	BLEU	0.12
style	PINC	0.41
style	Cosine	0.37
style	LM	0.45
style	Logistic regression	0.46

Table 4.2: Correlation between human judgments and automatic evaluation metrics when paraphrasing Shakespeare's plays into modern prose.

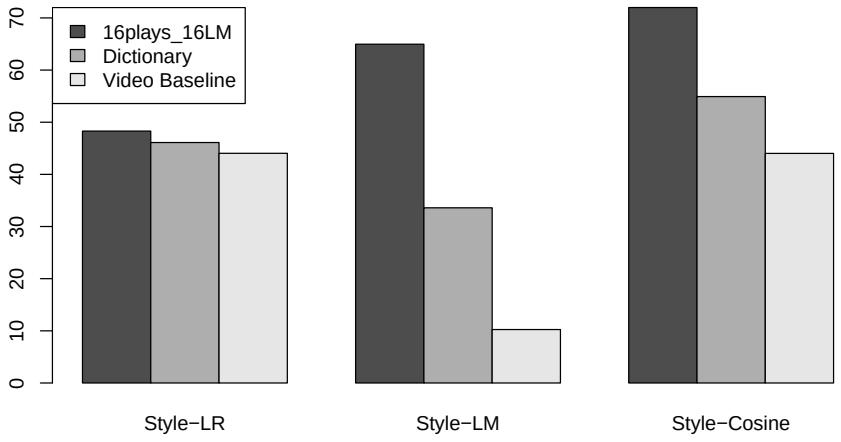

Figure 4.2: Automatic evaluation of paraphrasing Shakespeare's plays into modern English

4.5 Conclusions

We introduced three new metrics for evaluating writing style, one based on cosine similarity, one based on language models, and the third based on logistic regression. We measured correlation between automatic metrics and human judgments, and showed that our new metrics have better correlation with human judgments than existing metrics in the context of our task. While this evaluation is limited to one specific style of writing, we are optimistic that these or similar metrics will also perform well when evaluating paraphrase systems targeting other writing styles.

Chapter 5

Automatically Gathering and Generating Paraphrases from Twitter

Encouraged by our earlier success on paraphrasing historic English, we turn to the most up-to-date language variation, the Internet language. With hundreds of millions of users freely publishing their own content, social media services (e.g. Facebook, Twitter) provide a massive amount of valuable information as well as a vast diversity of linguistic expression, including the newly invented words and phrasing.

Learning paraphrases from Twitter posts (i.e. tweets) could be especially beneficial. First, the high level of information redundancy in Twitter provides a good opportunity to collect many different expressions. Second, tweets contain many kinds of paraphrases not available elsewhere including typos, abbreviations, ungrammatical expressions and slang, which can be particularly valuable for many applications, such as phrase-based text normalization (Kaufmann and Kalita, 2010) and correction of writing mistakes (Gamon et al., 2008), given the difficulty of acquiring annotated data. Paraphrase models that

are derived from microblog data could be useful to improve other NLP tasks on noisy user-generated text and help users to interpret a large range of up-to-date abbreviations (e.g. dlt → Doritos Locos Taco) and native expressions (e.g. oh my god → {oh my goodness | oh my gosh | oh my gawd | oh my jesus}) etc.

We present the first investigation into automatically collecting a large paraphrase corpus of tweets, which can be used for building paraphrase systems adapted to Twitter using techniques from statistical machine translation (SMT) (Xu et al., 2013c). We show experimental results demonstrating the benefits of an in-domain[1] parallel corpus when paraphrasing tweets. In addition, our paraphrase models can be applied to the task of normalizing noisy text where we show improvements over the state-of-the-art.

Relevant previous work has extracted sentence-level paraphrases from news corpora (Barzilay and Lee, 2003; Dolan et al., 2004; Quirk et al., 2004). Paraphrases gathered from noisy user-generated text on Twitter have unique characteristics which make this comparable corpus a valuable new resource for mining sentence-level paraphrases. Twitter also has much less context than news articles and much more diverse content, thus posing new challenges to control the errors in mining paraphrases while retaining the desired superficial dissimilarity.

5.1 Gathering A Parallel Tweet Corpus

There is a huge amount of redundant information on Twitter. When significant events take place in the world, many people go to Twitter to share, comment and dis-

1. In this chapter, the data derived from Twitter is considered in-domain; while other datasets, e.g. news and video descriptions, are out-of-domain.

cuss them. Among tweets on the same topic, many will convey similar meaning using widely divergent expressions. Whereas researchers have exploited multiple news reports about the same event for paraphrase acquisition (Dolan et al., 2004), Twitter contains more variety in terms of both language forms and types of events, and requires different treatment due to its unique characteristics.

As described in §5.1.1, our approach first identifies tweets which refer to the same popular event as those which mention a unique named entity and date, then aligns tweets within each event to construct a parallel corpus. To generate paraphrases, we apply a typical phrase-based statistical MT pipeline, performing word alignment on the parallel data using GIZA++ (Och and Ney, 2003), then extracting phrase pairs and performing decoding uses Moses (Koehn et al., 2007).

5.1.1 Extracting Events from Tweets

As a first step towards extracting paraphrases from popular events discussed on Twitter, we need a way to identify Tweets which mention the same event. To do this we follow previous work by Ritter et al. (2012), extracting named entities and resolving temporal expressions (for example "tomorrow" or "on Wednesday"). Because tweets are compact and self-contained, those which mention the same named entity and date are likely to reference the same event. We also employ a statistical significance test to measure strength of association between each named entity and date, and thereby identify important events discussed widely among users with a specific focus, such as the release of a new iPhone, as opposed to individual users discussing everyday events involving their phones. By gathering tweets based on popular real-world events, we can efficiently

extract pairwise paraphrases within a small group of closely related tweets, rather than exploring every pair of tweets in a large corpus. By discarding frequent but insignificant events, such as "I like my iPhone" and "I like broke my iPhone", we can reduce noise and encourage diversity of paraphrases by requiring less lexical overlap. Example events identified using this procedure are presented in Table 5.1.

Entity/Date	Example Tweets
Obama 11/6/2012	Vote for Obama on November 6th!
	OBAMA is #winning his 2nd term on November 6th 2012.
	November 6th we will re-elect Obama!!
James Bond 11/9/2012	Bought movie tickets to see James Bond tomorrow. I'm a big #007 fan!
	Who wants to go with me and see that new James Bond movie tomorrow?
	I wanna go see James Bond tomorrow
North Korea 12/29/2012	North Korea Announces December 29 Launch Date for Rocket
	Pyongyang reschedules launch to December 29 due to 'technical deficiency'
	North Korea to extend rocket launch period to December 29

Table 5.1: Example tweets taken from automatically identified significant events extracted from Twitter. Because many users express similar information when mentioning these events, there are many opportunities for paraphrase.

5.1.2 Extracting Paraphrases Within Events

Twitter users are likely to express the same meaning in relation to an important event, however not every pair of tweets mentioning the same event will have the same meaning. People may have opposite opinions and complicated events such as presidential elections

can have many aspects. To build a useful monolingual paraphrase corpus, we need some additional filtering to prevent unrelated sentence pairs.

If two tweets mention the same event and also share many words in common, they are very likely to be paraphrases. We use the Jaccard distance metric (Jaccard, 1912) to identify pairs of sentences within an event that are similar at the lexical level. Since tweets are extremely short with little context and include a broad range of topics, using only surface similarity is prone to unrelated sentence pairs. The average sentence length is only 11.9 words in our Twitter corpus, compared to 18.6 words in newswire (Dolan et al., 2004) which also contains additional document-level information. Even after filtering tweets with both their event cluster and lexical overlap, some unrelated sentence pairs remain in the parallel corpus. For example, names of two separate music venues in the same city might be mismatched together if they happen to have concerts on the same night that people tweeted using a canonical phrasing like "I am going to a concert at _____ in Austin tonight".

In addition to filtering out weakly associated sentence pairs, we also found it beneficial to perform additional filtering on the learned phrase translation table. We prune out unlikely phrase pairs using a technique proposed by Johnson et al. (2007) with their recommended setting, which is based on the significance testing of phrase pair co-occurrence in the parallel corpus (Moore, 2004). Based on the fact that each phrase is naturally equivalent in semantics to itself, we further prevent unreasonable translations by adding additional entries to the phrase table to ensure every phrase has an option to remain unchanged during paraphrasing.

5.2 Paraphrasing Tweets for Normalization

Paraphrase models built from grammatical text are not appropriate for the task of normalizing noisy text. However, the unique characteristics of the Twitter data allow our paraphrase models to include both normal and noisy language and consequently translate between them. Our models have a tendency to normalize because correct spellings and grammar appear more frequently than incorrect ones,[2] but there is still danger of introducing noise. For the purposes of normalization, we therefore biased our models using a language model built using text taken from the New York Times which is used to represent grammatical English.

Previous work on microblog normalization is mostly limited to word-level adaptation or out-of-domain annotated data. Our phrase-based models fill the gap left by previous studies by exploiting a large, automatically curated, in-domain paraphrase corpus.

Lexical normalization (Han and Baldwin, 2011) only considers transforming an out-of-vocabulary (OOV) word to its standard form, i.e. in-vocabulary (IV) word. Beyond word-to-word conversions, our phrase-based model is also able to handle the following types of errors without requiring any annotated data:

Error type	**Ill form**	**Standard form**
1-to-many	everytime	every time
incorrect IVs	can't want for	can't wait for
grammar	I'm going a movie	I'm going to a movie
ambiguities	4	4 / 4th / for / four

2. Even though misspellings and grammatical errors are quite common, there is much more variety and less agreement.

Kaufmann and Kalita (2010) explored machine translation techniques for the normalization task using an SMS corpus which was manually annotated with grammatical paraphrases. Microblogs, however, contain a much broader range of content than SMS and have no in-domain annotated data available. In addition, the ability to gather paraphrases *automatically* opens up the possibility to build normalization models from orders of magnitude more data, and also to produce up-to-date normalization models which capture new abbreviations and slang as they are invented.

5.3 Experiments

We evaluate our system and several baselines at the task of paraphrasing Tweets using previously developed automatic evaluation metrics which have been shown to have high correlation with human judgments (Chen and Dolan, 2011). In addition, because no previous work has evaluated these metrics in the context of noisy Twitter data, we perform a human evaluation in which annotators are asked to choose which system generates the best paraphrase. Finally we evaluate our phrase-based normalization system against a state-of-the-art word-based normalizer developed for Twitter (Han et al., 2012).

5.3.1 Paraphrasing Tweets

5.3.1.1 Data

Our paraphrase dataset is distilled from a large corpus of tweets gathered over a one-year period spanning November 2011 to October 2012 using the Twitter Streaming API. Following Ritter et al. (2012), we grouped together all tweets which mention the same

Input	Output
Hostess is going outta biz	hostess is going out of business
REPUBLICAN IMMIGRATION REFORM IS A THING NOW	gop imigration law is a thing now
Freedom Writers will always be one of my fav movies	freedom writers will forever be one of my favorite movies
sources confirm that Phil Jackson has cancelled all weekend plans and upcoming guest appearances, will meet with LAL front office	source confirms that phil jackson has canceled all weekend plans , upcomin guest appearances and will meet with lakers front office

Table 5.2: Example paraphrases generated by our system on the test data.

named entity (recognized using a Twitter specific named entity tagger[3]) and a reference to the same unique calendar date (resolved using a temporal expression processor (Mani and Wilson, 2000)). Then we applied a statistical significance test, the G^2 test (Moore, 2004), to rank the events, which considers the corpus frequency of the named entity, the number of times the date has been mentioned, and the number of tweets which mention both together. Altogether we collected more than 3 million tweets from the 50 top events of each day according to the p-value from the statistical test, with an average of 229 tweets per event cluster.

Each of these tweets was passed through a Twitter tokenizer[4] and a simple sentence splitter, which also removes emoticons, URLs, usernames and most of the hashtags and usernames. Hashtags and usernames that were in the middle of sentences and might be part of the text were kept. Within each event cluster, redundant and short sentences (less than 3 words) were filtered out, and the remaining sentences were paired together if

3. https://github.com/aritter/twitter_nlp
4. https://github.com/brendano/tweetmotif

their Jaccard similarity was not less than 0.5. This resulted in a parallel corpus consisting of 4,008,946 sentence pairs with 800,728 unique sentences.

We then trained paraphrase models by applying a typical phrase-based statistical MT pipeline on the parallel data, which uses GIZA++ for word alignment and Moses for extracting phrase pairs, training and decoding. We use a language model trained on the 3 million collected tweets in the decoding process. The parameters are tuned over a small set of development data.

Sentence alignment in comparable corpora is more difficult than between direct translations (Moore, 2002), and Twitter's noisy style, short context and broad range of content present additional complications. Our automatically constructed parallel corpus contains some proportion of unrelated sentence pairs and therefore does result in some unreasonable paraphrases. We reduce this noise by pruning out unlikely phrase pairs as suggested for statistical MT by Johnson et al. (2007). We further prevent unreasonable translations by adding additional entries to the phrase table to ensure every phrase has an option to remain unchanged during paraphrasing and normalization. Without these noise reduction steps, our system will produce paraphrases with serious errors (e.g. change a person's last name) for 100 out of 200 test tweets in the evaluation in §5.3.1.5.

In the meantime, it is also important to promote lexical dissimilarity in the paraphrase task. Following Ritter et al. (2011a) we add a lexical similarity penalty to each phrase pair in our system, in addition to the four basic components (translation model, distortion model, language model and word penalty) in SMT.

5.3.1.2 Evaluation Details

The beauty of the lexical similarity penalty is that it gives control over the degree of paraphrasing by adjusting its weight versus the other components. Thus we can plot a BLEU-PINC curve to express the tradeoff between semantic adequacy and lexical dissimilarity with the input, where BLUE (Papineni et al., 2002) and PINC (Chen and Dolan, 2011) are previously proposed automatic evaluation metrics to measure respectively the two criteria of paraphrase quality.

To compute these automatic evaluation metrics, we manually prepared a dataset of gold paraphrases by tracking the trending topics on Twitter[5] and gathering groups of paraphrases in November 2012. In total 20 sets of sentences were collected and each set contains 5 different sentences that express the same meaning. Each sentence is used once as input while the other 4 sentences in the same set serve as reference translations for automatic evaluation of semantic adequacy using BLEU.

5.3.1.3 Baselines

We consider two state-of-the-art paraphrase systems as baselines, both of which are trained on parallel corpora of aligned sentences. The first one is trained on a large-scale corpus gathered by asking users of Amazon's Mechanical Turk Service (Snow et al., 2008) to write a one-sentence description of a short video clip (Chen and Dolan, 2011). We combined a phrase table and distortion table extracted from this parallel corpus with the same Twitter language model, applying the Moses decoder to generate paraphrases. The additional noise removal steps described in §5.3.1.1 were found helpful for this

5. https://support.twitter.com/articles/101125-faqs-about-twitter-s-trends

model during development and were therefore applied. The second baseline uses the Microsoft Research paraphrase tables that are automatically extracted from news articles in combination with the Twitter language model.[6]

5.3.1.4 Results

Figure 5.1 compares our system against both baselines, varying the lexical similarity penalty for each system to generate BLEU-PINC curves. Our system trained on automatically gathered in-domain Twitter paraphrases achieves higher BLEU at equivalent PINC for the entire length of the curves. Table 5.2 shows some sample outputs of our system on real Twitter data.

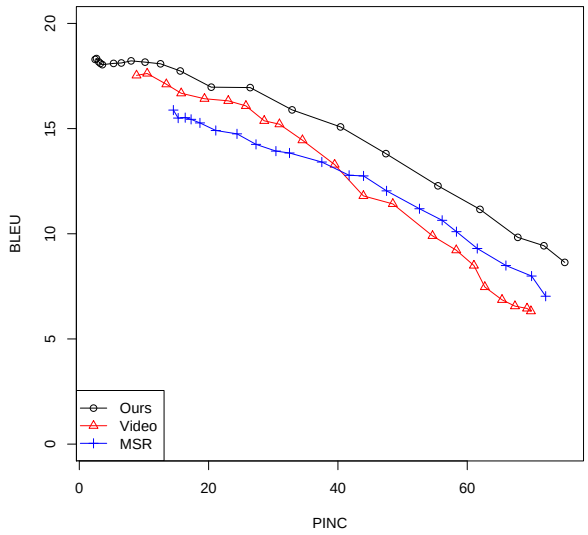

Figure 5.1: Results from automatic paraphrase evaluation. PINC measures n-gram dissimilarity from the source sentence, whereas BLEU roughly measures n-gram similarity to the reference paraphrases.

6. No distortion table or noisy removal process is applied because the parallel corpus is not available.

One novel feature of our approach, compared to previous work on paraphrasing, is that it captures many slang terms, acronyms, abbreviations and misspellings that are otherwise hard to learn. Several examples are shown in table 5.3. The rich semantic redundancy in Twitter help generate a large variety of typical paraphrases as well (see an example in table 5.4).

Input	**Top-ranked Outputs**
amped	pumped
lemme kno	let me know
bb	bigbang, big brother
snl	nbcsnl, saturday night live
apply 4 tix	apply for tickets, ask for tickets, applying for tickets
the boys	one direction (a band, whose members are often referred as "the boys"), they, the boy, the gys, the lads, my boys, the direction (can be used to refer to the band "one direction"), the onedirection, our boys, our guys
oh my god	oh my gosh, omfg, thank the lord, omg, oh my lord, thank you god, oh my jesus, oh god
can't wait	cant wait, cant wait, cannot wait, i cannot wait, so excited, cnt wait, i have to wait, i can'wait, ready, so ready, so pumped, seriously can'wait, really can't wait

Table 5.3: Example paraphrases of noisy phrases and slang commonly found on Twitter

5.3.1.5 Human Evaluation

In addition to automatic evaluation, we also performed a human evaluation in which annotators were asked to pick which system generated the best paraphrase. We used the same dataset of 200 tweets gathered for the automatic evaluation and generated paraphrases using the 3 systems in Figure 5.1 with the highest BLEU which achieve a PINC of at least 40. The human annotators were then asked to pick which of the 3

Input	Top-ranked Outputs
who want to get a beer	wants to get a beer, so who wants to get a beer, who wants to go get a beer, who wants to get beer, who want to get a beer, trying to get a beer, who wants to buy a beer, who wants to get a drink, who wants to get a rootbeer, who trying to get a beer, who wants to have a beer, who wants to order a beer, i want to get a beer, who wants to get me a beer, who else wants to get a beer, who wants to win a beer, anyone wants to get a beer, who wanted to get a beer, who wants to a beer, someone to get a beer, who wants to receive a beer, someone wants to get a beer

Table 5.4: Example paraphrases of a given sentence "who want to get a beer"

systems generated the best paraphrase using the criteria that it should be both different from the original and also capture as much of the original meaning as possible. The annotators were asked to abstain from picking one as the best in cases where there were no changes to the input, or where the resulting paraphrases totally lost the meaning.

Figure 5.2 displays the number of times each annotator picked each system's output as the best. Annotator 2 was somewhat more conservative than annotator 1, choosing to abstain more frequently and leading to lower overall frequencies, however in both cases we see a clear advantage from paraphrasing using in-domain models. As a measure of inter-rater agreement, we computed Cohen's Kappa between the annotators' judgment as to whether the Twitter-trained system's output the best. The value of Cohen's Kappa in this case was 0.525.

5.3.2 Phrase-Based Normalization

Because Twitter contains both normal and noisy language, with appropriate tuning, our models have the capability to translate between these two styles, e.g. paraphras-

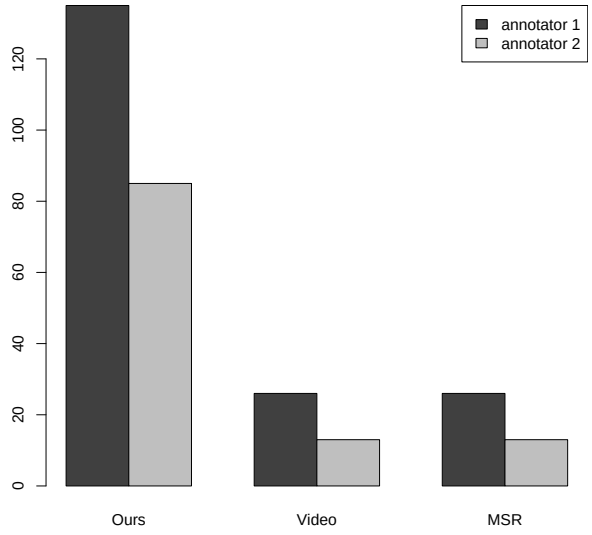

Figure 5.2: Results of human evaluation on paraphrasing Tweets.

ing into noisy style or normalizing into standard language. Here we demonstrate its capability to normalize tweets at the sentence-level.

5.3.2.1 Baselines

Much effort has been devoted recently to developing normalization dictionaries for Microblogs. One of the most competitive dictionaries available today is HB-dict+GHM-dict+S-dict used by Han et al. (2012), which combines a manually-constructed Internet slang dictionary , a small (Gouws et al., 2011) and a large automatically-derived dictionary based on distributional and string similarity. We evaluate two baselines using this large dictionary consisting of 41181 words; following Han et al. (2012), one is a simple dictionary look up. The other baseline uses the machinery of statistical machine translation, using this dictionary as a phrase table in combination with Twitter and NYT

language models.

5.3.2.2 System Details

Our base normalization system is the same as the paraphrase model described in §5.3.1.1, except that the distortion model is turned off to exclude reordering. We tuned the system towards correct spelling and grammar by adding a language model built from all New York Times articles written in 2008. We also filtered out the phrase pairs which map from in-vocabulary to out-of-vocabulary words. In addition, we integrated the dictionaries by linear combination to increase the coverage of the phrase-based SMT model (Bisazza et al., 2011).

5.3.2.3 Evaluation Details

We adopt the normalization dataset developed by Han and Baldwin (2011), which includes 549 random sampled English tweets. This dataset was initially annotated for the token-level normalization task: only words, that are outside a predefined vocabulary and determined by annotators as ill-formed, are one-to-one mapped to their standard forms. We augmented with sentence-level annotations, as shown in Table 5.5, to better fit the real practices of noisy text normalization.

5.3.2.4 Results

Normalization results are presented in figure 5.6. Using only our phrase table extracted from Twitter events we achieve poorer performance than the state-of-the-art dictionary baseline, however we find that by combining the normalization dictionary of

Original Tweet	oh alright alright i'll hit u up later cause im bout to get off
Token-level Annotation (Han and Baldwin, 2011)	oh alright alright i'll hit you up later cause i'm bout to get off
Sentence-level Annotation (This Work)	oh alright alright i'll hit you up later because i'm about to get off
Original Tweet	hehe its ok ! tty tomorrow xx byeee
Token-level Annotation (Han and Baldwin, 2011)	hehe its ok ! tty tomorrow xx bye
Sentence-level Annotation (This Work)	hehe it is ok ! talk to you tomorrow xx bye

Table 5.5: Examples from the Twitter normalization dataset

	BLEU	**PINC**
No-Change	60.00	0.0
SMT+TwitterLM	62.54	5.78
SMT+TwitterNYTLM	65.72	9.23
Dictionary	75.07	22.10
Dicionary+TwitterNYTLM	75.12	20.26
SMT+Dictionary+TwitterNYTLM	77.44	25.33

Table 5.6: Normalization performance

Han et al. (2012) with our automatically constructed phrase-table we are able to combine the high coverage of the normalization dictionary with the ability to perform phrase-level normalizations (e.g. "outta" → "out of" and examples in §5.2) achieving both higher PINC and BLEU than the systems which rely exclusively on word-level mappings. Our phrase table also contains many words that are not covered by the dictionary (e.g. "pts" → "points", "noms" → "nominations").

5.4 Conclusions

We have presented the first approach to gathering parallel monolingual text from Twitter, and built the first in-domain models for paraphrasing tweets. By paraphrasing using models trained on in-domain data we showed significant performance improvements over state-of-the-art out-of-domain paraphrase systems as demonstrated through automatic and human evaluations. We showed that because tweets include both normal and noisy language, paraphrase systems built from Twitter can be fruitfully applied to the task of normalizing noisy text, covering phrase-based normalizations not handled by previous dictionary-based normalization systems.

Chapter 6

Twitter Paraphrase Collection via Crowdsourcing

We have demonstrated the feasibility and utility of paraphrasing two distinct variations of English, i.e. the historic Shakespearean and the still-evolving Internet language. We learned paraphrase models in both domains by applying statistical phrase-based machine translation techniques on parallel text. Yet these two domains are very different in the way that parallel data can be derived.

William Shakespeare, as one of the most productive and most studied writers who ever lived, has developed a considerable amount of text in a consistent style, which are also translated into modern English by experts. In contrast, the Internet texts are spontaneously generated by hundreds of millions of individual users, which makes it more difficult to mine parallel texts. In the previous chapter we presented a solution using event extraction techniques and filtering mechanisms to find parallel sentences in Twitter. While this fully automated procedure can produce paraphrase corpus in large quantity,

it is limited by the coverage and accuracy of the event extraction components. Here we investigate alternative ways of collecting paraphrase data from Twitter by leveraging crowdsourcing.

In this chapter, we present a crowdsourcing approach to obtain good quality datasets that contain representative examples of both paraphrases and non-paraphrases in Twitter. Several data filtering and selection methods are experimented with to improve the efficiency of data collection by crowdsourcing. We also demonstrate the utility of these data for training and evaluating paraphrase identification systems in the social media domain, and show some useful insights with regard to the opportunities and challenges of paraphrases in colloquial language for future study.

6.1 Raw Data from Twitter

We crawled the trending topics and their tweets from Twitter. Trends on Twitter are determined by an undisclosed algorithm which identifies topics that are immediately popular, rather than topics that have been popular for a while or on a daily basis[1]. We then split each tweet into sentences, and collapsed very redundant sentences under each topic (e.g. sentences which only differ by punctuation).

6.2 Task Design on Mechanical Turk

We convert the task of finding sentential paraphrases from the Twitter trends into a simple and efficient task on Amazon Mechanical Turk. We showed the annotator

1. https://support.twitter.com/articles/101125-faq-about-trends-on-twitter

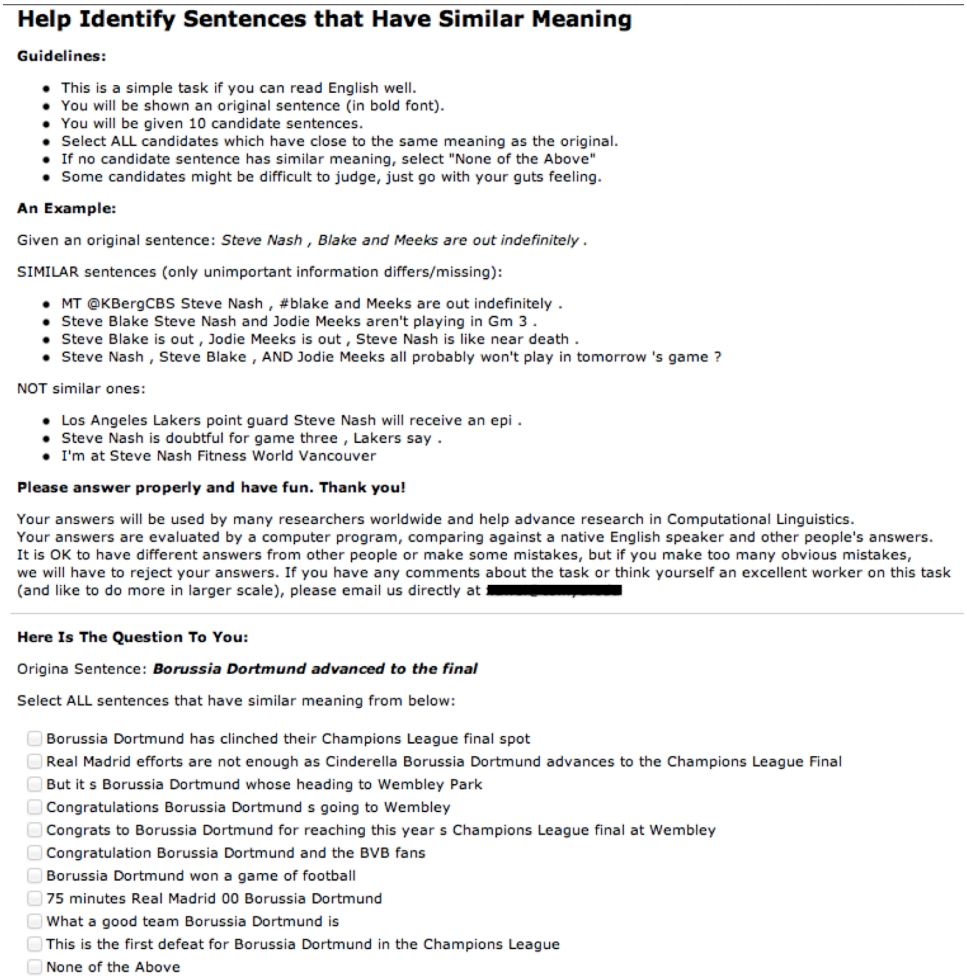

Figure 6.1: A screenshot of our annotation task as it was deployed on Amazon's Mechanical Turk

an **original** sentence, then ask them to pick sentences with the same meaning from 10 **candidate** sentences. For each of such 1vs10 questions, we obtained binary judgements of 10 sentence pairs from 5 different annotators, paying each annotator $0.02 per question. A screen shot of our annotation task is shown in Figure 6.1. On average, each question takes an annotator about 30 – 45 seconds to answer.

6.3 Annotation Quality

We evaluate the quality of annotators by computing each of their Cohen's Kappa score (Artstein and Poesio, 2008) against the majority vote among the other 4 annotations, excluding those cases of a tie. We block those annotators of poor quality on Amazon's Mechanical Turk[2]. Despite the difficulty to precisely define **paraphrase** (Ho et al., 2012) or **similar meaning**, the inter-rater reliability of Fleiss' Kappa (Joseph, 1971) among annotators is about 0.40, indicating "moderate agreement" (Landis and Koch, 1977).

To assess the reliability of the paraphrase annotation procedure, we conducted an agreement study with an expert annotator, using 971 randomly chosen sentence pairs across 40 trending topics. In our expert annotation, we adapted a fine-grained Likert scale ranging from 0 to 5 to measure the semantic similarity between sentences, which is defined by Agirre et al. (2012) as follows:

5 - Completely equivalent, as they mean the same thing;

4 - Mostly equivalent, but some unimportant details differ;

3 - Roughly equivalent, but some important information differs/missing.

2 - Not equivalent, but share some details;

1 - Not equivalent, but are on the same topic;

0 - On different topics.

Note that the annotation we obtained through crowdsourcing is also on a scale from 0 to 5, in terms of the number of annotators that identify a given sentence pair as having

2. During the entire annotation process, we detected two problematic workers.

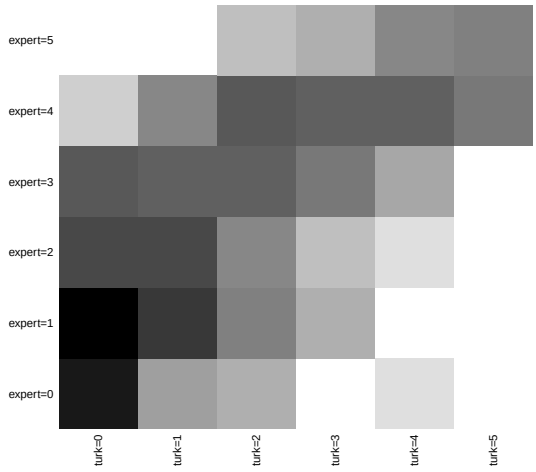

Figure 6.2: A heat-map showing overlap between expert and crowdsourcing annotation. Note that the two annotation scales are defined differently.

similar meaning. Although the two scales of expert and crowdsourcing annotation are defined differently, their Pearson correlation coefficient reaches 0.735 (two-tailed significance 0.001). Figure 6.2 shows a heat-map representing the detailed overlap between the two annotations. The intensity along the diagonal reflects good reliability of crowdsourcing workers for this particular task; and the shift above the diagonal presents the difference between the two annotation metrics, e.g. 0,1,2 in expert annotation all mean non-paraphrase, while for crowdsourcing the numbers indicate how many annotators out of 5 picked the sentence pair as paraphrases.

6.4 Selecting Sentences for Efficient Annotation

Since Twitter users are free to talk about anything regarding any topic, a random pair of sentences about the same topic has a low chance to express the same meaning. The small proportion of paraphrases causes two problems: 1) it gets too expensive to obtain paraphrases via manual annotation; 2) non-expert annotators tend to loosen the criteria and more likely make false positive errors. We need some filtering mechanism that promotes sentences which are more likely to be paraphrases to annotators, while preserving diversity and representativeness.

6.4.1 Automatic Summarization Inspired Filtering

Our inspiration comes from a typical problem in extractive summarization of multiple documents (Li et al., 2006; Nenkova and Vanderwende, 2005) or tweets (Xu et al., 2013a), that the salient sentences are likely to share redundant information. We utilize the scoring method used in SumBasic (Nenkova and Vanderwende, 2005; Vanderwende et al., 2007), a simple but strong summarization system, to sort sentences by salience. For each topic, SumBasic first computes the probability of each word by simply counting its frequency in all the sentences. Each sentence is scored as the average of the probabilities of the words in it. Unlike SumBasic, we do not take further steps to eliminate redundancy. This scoring schema prefers sentences with more frequent words, which means less specific and more common information and thus higher possibility of paraphrases. We select the **original** sentence randomly from the most (top 10%) salient sentences and **candidate** sentences from medium (top 50%) salient sentences to present

to the annotators [3].

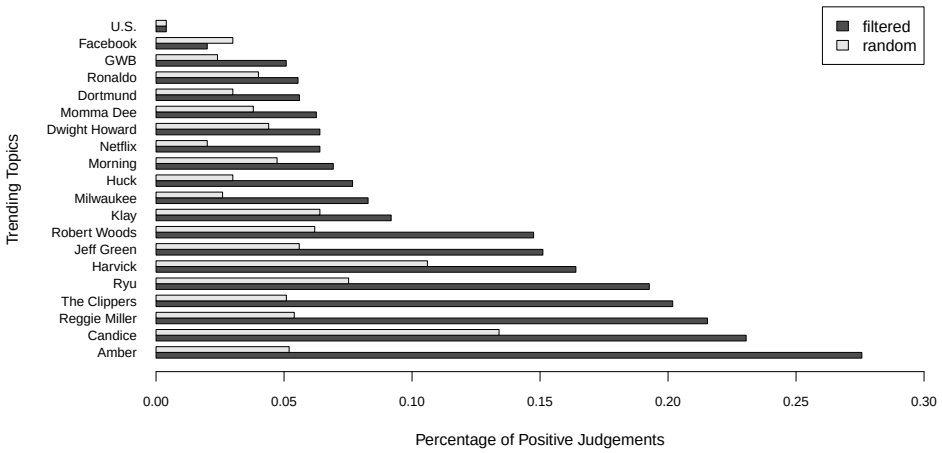

Figure 6.3: The proportion of paraphrases (percentage of positive votes from annotators) vary across different topics

6.4.2 Filtering vs. Random Selecting Experiment

In a trial experiment of 20 topics, with filtering, the number of sentence pairs labeled as paraphrases by the majority of annotators (>= 3 annotators out of 5) is twice more than random selection, increasing from 152 to 329 out of 2000 pairs (Figure 6.3 and Figure 6.4). We also use PINC (Chen and Dolan, 2011) to measure the quality of paraphrases collected according to n-gram dissimilarity. As shown in Figure 6.5, both the filtered and random paraphrases collected from Twitter have very high PINC scores, exhibiting significant rewording.

3. Due to limited budget and low proportion of paraphrases, we are unable to collect enough (estimated 4000 questions, $0.125/question) data to optimize the two thresholds. (*need to revise this sentence)

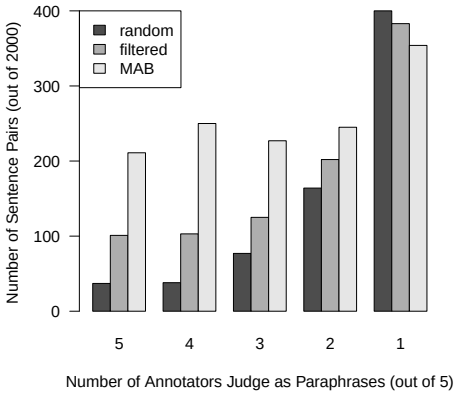

Figure 6.4: Numbers of paraphrases collected by different methods

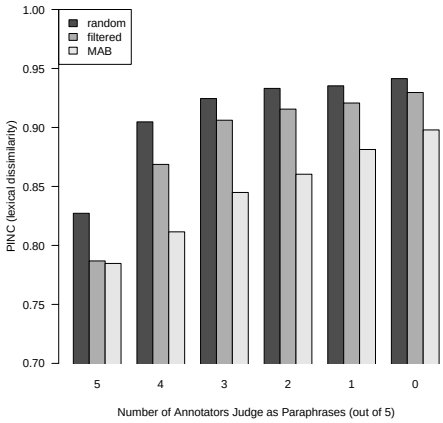

Figure 6.5: PINC scores of paraphrases collected

6.5 Selecting Topics for Efficient Annotation

As shown in Figure 6.3, the number of paraphrases varies greatly from topic to topic and thus the chance to encounter paraphrases during annotation. It arises the exploration versus exploitation dilemma for annotation strategy, i.e. the search for a balance between exploring the topics to find a profitable topic to annotate while annotating the empirically best topic (most paraphrases) as often as possible. This can be viewed as a Multi-Armed Bandit (MAB) (Robbins, 1985) problem.

6.5.1 Effective Crowdsourcing using Multi-Armed Bandits

The task of selecting topics in Twitter for more efficient paraphrase annotation can be formalized as a Multi-Armed Bandit problem with the following constraints:

- Infinite Arms: unlimited trending topics on Twitter;
- Budgeted (Finite Horizon): limited budget, uniform cost for each pull;
- Bounded: only a small number of paraphrases are needed from each topic for diversity.

The MAB model consists of a slot machine with unlimited numbers of arms (topics), denoted by $a_1, a_2, \ldots, a_\infty$. At each time step t, an agent chooses a bag $S(t)$ of arms to pull (annotate). By pulling arm a_i, the agent has to pay a uniform pulling cost and receive a non-negative reward (number of paraphrases) drawn from a distribution associated with that specific arm. The agent has a cost budget to pull at most B times during the entire annotation process. We also have the constraint that the agent cannot pull each arm more than l times in total. We use μ_i to denote the mean value of the rewards that the agent receives from pulling arm a_i. The agent has no initial knowledge of the reward of each arm and has to pull the arms to learn about them. The goal of the agent is to maximize the sum of rewards it earns from pulling the arms of the machine, with respect to the budget B.

6.5.2 Bounded ϵ-first Algorithm for MAB with Infinite Arms

Given μ_i are unknown *a priori*, the agent has to *explore* these values by pulling a particular arm in order to estimate its expected reward value. In contrast, the agent also need to *exploit* the best arm (based on previous observation) to maximize the total reward. The most related work is by Tran-Thanh et al. (2012) presenting a greedy strategy, called **bounded ϵ-first** for their budgeted bounded MAB problem, which has

a small number of arms. We modified their algorithm to accommodate our task with an infinite number of arms:

Our strategy consists of two phases. In the first exploration phase t_0, we dedicate an ϵ portion of budget B to estimate the expected reward values of the arms. We randomly select $\lceil \frac{\epsilon B}{m} \rceil$ arms and pull each m times (except the last arm). After this exploration phase, we update the estimated reward $\hat{\mu}_i(t_0)$ for each arm pulled by taking the average of received reward (percentage of paraphrases identified by annotators) from arm a_i. In the second exploitation phase t_1, our goal is to maximize the reward by selecting from the explored arms to pull:

$$\begin{aligned} \max \quad & \sum_{j \in \{i | r_i(t_0) > 0\}} \hat{\mu}_i(t_0) r_i(t_1) \\ \text{s.t.} \quad & \sum_{j \in \{i | r_i(t_1) > 0\}} r_i(t_0) \leq (1-\epsilon)B, \forall i : 0 \leq r_i(t_1) \leq l - r_i(t_0). \end{aligned} \tag{6.1}$$

where $r_i(t_1)$ denotes the times of arm a_i is chosen to be pulled at the exploitation phase t_1. We exploit a simple yet efficient approximation method (Kellerer et al., 2004) to solve this problem. We sort all the arms in an non-decreasing order according to $\hat{\mu}_i(t_0)$, and we sequentially pull $\lceil \frac{(1-\epsilon)B}{l-m} \rceil$ arms that have the highest estimated rewards until reaching l times including the pulls in the exploration phase (except the last arm).

6.5.3 Simulation and Real-world Experiments

To analyze the behavior of the proposed algorithm and optimize the parameters, we artificially expand a small amount of real annotation data over 40 topics by duplicating, and then conduct simulation experiments on the enlarged dataset. The experiments

with various settings showed the best value of m to be 1, indicating the importance of exploration in infinite arm situation and low false positive rate in estimating the number of paraphrases for each topic by only one initial test. Figure 6.6 shows the algorithm performance, having the limit of each topic l set to 10 questions (100 sentence pairs) for annotators and the total budget set to 1500 questions. While the existing few empirical studies on multi-armed bandit algorithms (Tran-Thanh et al., 2012; Vermorel and Mohri, 2005) report best performances with small epsilons at about 0.05 - 0.15, our infinite arm problem prefers higher epsilon around 0.35-0.55 and more exploration.

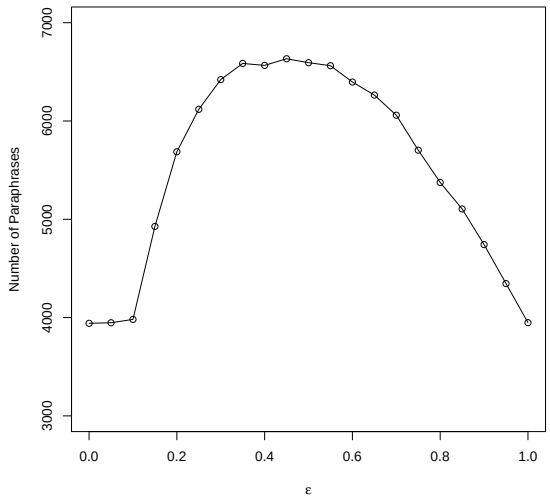

Figure 6.6: Simulation analysis of Bounded ϵ-first Algorithm for Infinite Arms

After the simulation, we apply the MAB algorithm to obtain real-world paraphrase data. We end up using $175 to explore 500 trending topics randomly by asking one question per topic, and then exploit 100 topics with nine questions for each topic, which has $\epsilon = 0.36$. Working on the same raw Twitter dataset, the number of paraphrases

(sentence pairs labeled as having similar meaning by the majority of annotators, >= 3 annotators out of 5) collected on Amazon Mechanical Turk is increased to 688 out of 2000 sentence pairs by using MAB together with filtering, compared to 329 by using filtering alone (Figure 6.4). The paraphrases collected using MAB have lower lexical dissimilarity, but still maintain PINC scores relatively high at over 0.75, as shown in Figure 6.5.

6.6 Utilizing the Collected Data for Paraphrase Identification in Twitter

To investigate the task of paraphrase detection in Twitter, we evaluate several existing supervised and unsupervised approaches utilizing the dataset we collected through crowdsourcing. As the preliminary experiment results will show, our Twitter paraphrase corpus is adequately reliable for training and testing paraphrase identification systems in the Twitter domain. However, the performance of prior approaches declines comparing to the traditional news domain. This shows challenges in paraphrasing colloquial language while great space for improvement and research potential.

6.6.1 Supervised Learning Approaches

We reimplement the logistic regression (LR) model by Das and Smith (2009) for paraphrase detection. It was based on lexical overlap features and reported to be on par with a strong state-of-the-art baseline by Wan et al. (2006) using SVM with dependency-based features. When trained and tested on the MSRP corpus (Dolan et al., 2004), Das

and Smith's LR model achieved 0.78 precision, 0.87 recall and 0.75 accuracy; while Wan et al.'s SVM model obtained 0.77, 0.90 and 0.75 respectively.

We also propose a new set of lexical overlap features based on n-gram Jaccard similarity and dissimilarity (referred as Basic Features) for the LR model, that are different from the original feature set used by Das and Smith (2009). We use the paraphrases collected via Mechanical Turk from the first week by various methods as training data, and 1000 sentence pairs sampled from second week as test data. The first week and second week data is collected one month apart. The experiment results are shown in table 6.1. The cases that are on the edge, those with an expert annotation of score 3 and an crowdsourcing annotation of 2 (2 positive votes out of 5 annotators), are discarded in both training and test set, resulting in slight difference in training data size.

Training (week-size)	Features	Test - Expert Annotation				Test - Crowdsourcing Annotation			
		P	R	F	A	P	R	F	A
1w-3381	Das & Smith (2009)	71.30	47.13	56.75	85.06	67.54	47.53	55.80	85.83
40 topics	§6.6.1 Basic Features	77.00	44.25	56.20	85.66	71.57	45.06	55.30	86.30
2w-3598 *	Das & Smith (2009)	71.43	40.23	51.47	84.23	62.00	38.27	47.33	83.97
40 topics	§6.6.1 Basic Features	78.57	37.93	51.16	84.95	71.59	38.89	50.40	85.60
1w-3617	Das & Smith (2009)	71.79	32.18	44.44	83.27	70.67	32.72	44.73	84.79
400 topics	§6.6.1 Basic Features	80.77	36.21	50.00	84.95	77.22	37.65	50.62	86.18
1w-3487	Das & Smith (2009)	60.00	55.17	57.49	83.03	56.52	56.17	56.35	83.62
bandit	§6.6.1 Basic Features	66.67	55.17	60.38	84.95	64.83	58.02	61.24	86.18
1w-12271	Das & Smith (2009)	65.77	56.32	60.68	84.83	63.16	59.26	61.15	85.83
bandit	§6.6.1 Basic Features	67.11	57.47	61.92	85.30	63.76	58.64	61.09	85.95

Table 6.1: Classification accuracy, positive-class precision, recall and F-measure of different systems, trained by different data and tested on same dataset using two annotations as golden labels (* train and test on the same data set by leave-one-topic cross-validation)

Some conclusions:

– Crowdsourcing annotation is good enough to be used as gold data for evaluation. ('Expert Annotation' vs. 'Crowdsourcing Annotation')

- Using data collected from the same time period as test data for training does not achieve better performance than using training data from a different time period, due to high-dimensional lexical *overlap* features. ('1w-3381 40 topics - different period' vs. '2w-3598 40 topics - same period')

- If size of training data is set to 4000 sentence pairs, using 40 topics x 100 pairs as training data is better than 400 topics x 10 pairs. ('1w-3381 40 topics' vs. '1w-3617 400 topics')

- Data collected using MAB is biased towards higher proportion of positive cases, and got higher recall but lower precision. ('bandit' vs. else)

- More data collected using MAB does not help, possibly because of the features are based on lexical similarity and dissimilarity between sentences. The size of training data may have a bigger impact if systems use surface string features etc. ('1w-3487 bandit' vs. '1w-12271')

6.6.2 Unsupervised Learning Approaches

Using the dataset we collected as a test bed, we also evaluate the capability of a state-of-the-art semantic similarity measure (Guo and Diab, 2012) on Twitter data. It is a latent semantic approach that is specially developed for short sentences/texts by modeling the semantic space of not only the words present but also absent in the sentences. The intuition is that, given the little context of the short sentence, it is helpful to also take the information that is missing from the sentence into consideration.

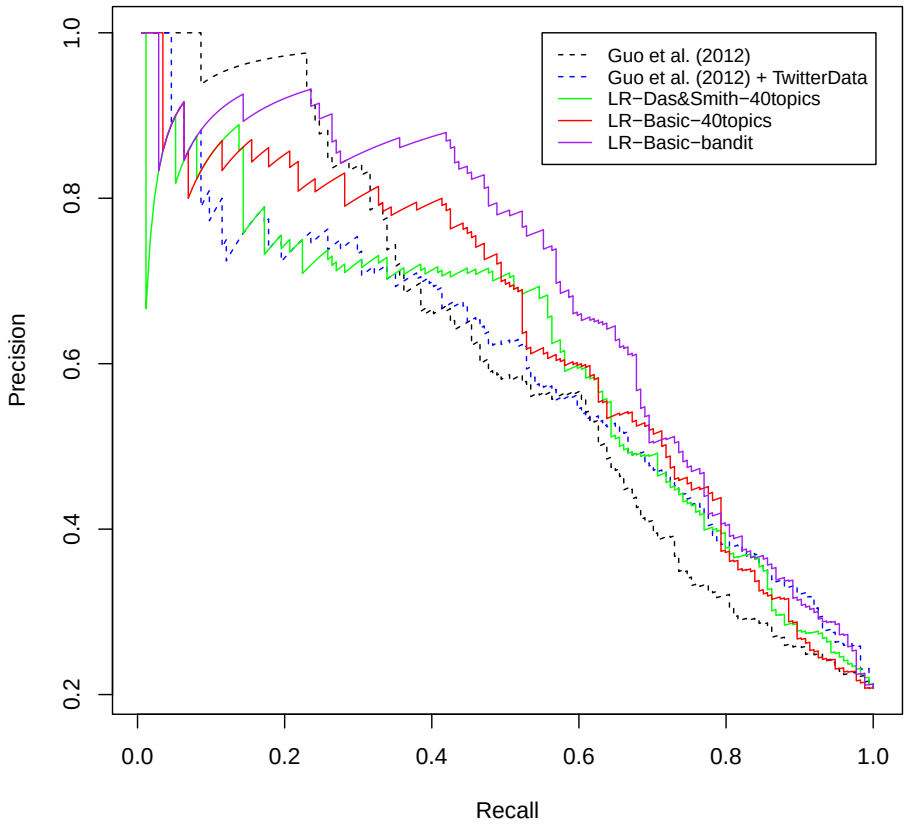

Figure 6.7: Precision-Recall curves comparing supervised and unsupervised approaches for paraphrase identification. The dashed plots represent unsupervised methods; while solid plots represent supervised methods.

We used the same evaluation setting as in §6.6.1 and evaluated against expert annotation. Two latent semantic models are compared: one is built on the original dataset used in (Guo and Diab, 2012) which includes WordNet (Fellbaum, 2010), OntoNotes (Hovy et al., 2006), Wiktionary[4] and Brown corpus (Francis and Kucera, 1979); while

4. http://www.wiktionary.org/

the other one also uses 1.6 million sentences from Twitter in addition. The results are shown in Figure 6.7. The unsupervised models achieve comparable but not better results comparing to supervised methods.

6.7 Conclusions

We have presented a successful approach to collecting a paraphrase corpus from Twitter by crowdsourcing efforts. In addition, we improved the efficiency of this data collection procedure by about 4 times by selecting better sentences and topics to annotate.

Using the paraphrase corpus annotated by crowd, we train and evaluate several prior approaches for paraphrase identification in the new Twitter domain. The preliminary experiment results look promising, but show a performance drop compared to the news domain. This motivates us to develop paraphrasing techniques that are more suitable to the colloquial language in the future.

Chapter 7

Future Work

My thesis focuses on a neglected but interesting and important problem in computational linguistics, paraphrasing across varieties of a language. Sitting in between two popular research fields, bilingual machine translation and monolingual paraphrasing, it introduces new challenges, but also new potentials to model language variation in many NLP applications. We investigated the challenges of data collection and experimented with different approaches, varying from expert translation, crowdsourcing annotation and fully automatic methods. We showed encouraging results and demonstrated the potential of modeling this special kind of paraphrase on Shakespearean and Internet texts, which goes beyond the scope of the prior study on paraphrasing in NLP. We also laid the foundation for future work by proposing evaluation schemas for this new task.

As this is the first systematic study of paraphrases in language variation, it presents many opportunities for future work:

7.1 Diversity-aware Automatic Paraphrase Identification for Twitter

It has been a long-standing problem to identify paraphrases with great lexical diversity (Chen and Dolan, 2011; Dolan et al., 2004; Shima and Mitamura, 2011), which motivated our work. While we have shown that Twitter is a great data source to solve this problem and automatic paraphrase identification is practicable, we plan to improve further Twitter-specific approaches that emphasize both diversity and precision. In contrast to the previous work, as discussed in §2.1, which focuses on precision and recall, we argue that given the huge volume of data available daily on Twitter, recall is not as crucial but the opportunity to obtain paraphrases with great lexical diversity is more important. To achieve this goal, we consider two possible directions that take advantage the unique characteristics of Twitter data:

- **Joint Word and Sentence Alignment**. Alignment is more difficult with language variations than bilingual data (Brown et al., 1990) or monolingual corpora (Barzilay and Elhadad, 2003; Nelken and Shieber, 2006). Besides spelling variations, we also face some complex cases in language change where alignment between large phrases is required. These difficult sentence pairs confuse the IBM word alignment models in GIZA++ used in our initial experiments. Since we do not have human annotated word alignment data required by the current phrase-based monolingual alignment (Yao et al., 2013), we plan to validate word and phrase mappings directly using exact significance tests (Mehta and Patel, 1997; Moore, 2004) on co-occurrences and reduce the sparsity problem by considering the transitivity of

paraphrases (e.g. "hotlanta" → "atl", "atl" → "atlanta"). We could perform word and sentence alignment jointly in a bootstrapping fashion, i.e. improve sentence alignment by using a lexicon, then expand the lexicon from the updated parallel sentences, and iterate. We also consider utilizing the parallel corpus annotated via crowdsourcing as distant supervision (Hoffmann et al., 2011; Mintz et al., 2009; Riedel et al., 2010; Ritter et al., 2013; Surdeanu et al., 2012; Xu et al., 2011b, 2013b) in training word aligners.

- **Integrating Cross-Tweet Information** Although a single tweet is often too short for analysis, many related tweets can provide a fuller context for what are paraphrases. We have experimented with this idea to locate paraphrases by a combination of event extraction and lexical overlapping in §5. We would like to extend it further to explicitly model topic-level information, e.g. significant words of the topic, as features in the machine learning approaches. In this way, we could consider all kinds of words or phrases in addition to name entities and temporal relations to anchor paraphrases.

7.2 Paraphrasing for Colloquial English

Language variation study is the major subject in sociolinguistic (Chambers and Schilling, 2013) and recently received some attention from computational linguistics. However, most computational linguists have been focusing on identifying the differences (Volkova et al., 2013) other than the equality between language variations, by which we stress semantic equivalence. We are particularly interested in the following

three subtopics for future work:

- **Language Identification of Colloquial English**. We would like to distinguish between normal English and colloquial style automatically, since the two are often mixed together in the real world. This would help adapting NLP tools to informal text more accurately. We can also alternate normal English to fit Twitter's specific style automatically for various usages. We plan to learn a language identifier by using various training data, such as annotated tweets using Amazon's Mechanical Turk, posts from news media accounts on Twitter and tweets containing URLs linking to news articles. A recent study on linking tweets to news (Guo et al., 2013) also gives us some insights to start with.

- **Idiom Detection in Twitter and Dictionary Construction**. The language used in social media is rich in idioms due to both of its informal nature and short length allowed. It would be very helpful for many NLP applications, if we could identify idioms in tweets and compile an idiom dictionary that connects synonyms. Although there has been much previous work on idiom detection (Birke and Sarkar, 2006; Li and Sporleder, 2009), not much has been done in social media domain. Encouraged by the positive report on idioms identification in Wiktionary (Muzny and Zettlemoyer, 2013), we are particularly interested in utilizing and extending existing idiom dictionaries with Twitter data. Besides Wiktionary (Zesch et al., 2008), another user-generated online dictionary, Urban Dictionary[1], is poorly organized though it contains useful information. We consider designing an automatic method to compile it into a more useful linguistic resource with better

1. http://www.urbandictionary.com/

quality. We also think it should be possible to extract explanations of idioms by finding redundant tweets of the idiomatic ones.

Bibliography

Agirre, Eneko, Diab, Mona, Cer, Daniel, and Gonzalez-Agirre, Aitor (2012). "Semeval-2012 task 6: A pilot on semantic textual similarity". In: *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 385–393.

Artstein, Ron and Poesio, Massimo (2008). "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4, pp. 555–596.

Bacchiani, M. and Roark, B. (2003). "Unsupervised language model adaptation". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE, pp. I–224.

Baldwin, Timothy, Cook, Paul, Lui, Marco, MacKinlay, Andrew, and Wang, Li (2013). "How Noisy Social Media Text, How Diffrnt Social Media Sources?" In: *Proceedings of the 6th International Joint Conference on Natural Language Processing*.

Bannard, Colin and Callison-Burch, Chris (2005). "Paraphrasing with Bilingual Parallel Corpora". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

Barzilay, Regina and Lee, Lillian (2003). "Learning to paraphrase: an unsupervised approach using multiple-sequence alignment". In: *Proceedings of the 2003 Conference of*

the North American Chapter of the Association for Computational Linguistics on Human Language Technology.

Barzilay, R. and McKeown, K.R. (2001). "Extracting paraphrases from a parallel corpus". In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 50–57.

Barzilay, R. and Elhadad, N. (2003). "Sentence alignment for monolingual comparable corpora". In: *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, pp. 25–32.

Birke, Julia and Sarkar, Anoop (2006). "A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language." In: *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*.

Bisazza, Arianna, Ruiz, Nick, and Federico, Marcello (2011). "Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation". In: *International Workshop on Spoken Language Translation*.

Wang, Ling, Dyer, Chris, Black, Alan W., and Trancoso, Isabel (2013). "Paraphrasing 4 Microblog Normalization". In: *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

Blitzer, J., McDonald, R., and Pereira, F. (2006). "Domain adaptation with structural correspondence learning". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 120–128.

Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., and Roossin, P.S. (1990). "A statistical approach to machine translation". In: *Computational linguistics* 16.2, pp. 79–85.

BIBLIOGRAPHY

Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., and Mercer, R.L. (1993). "The mathematics of statistical machine translation: Parameter estimation". In: *Computational linguistics* 19.2, pp. 263–311.

Burrows, Steven, Potthast, Martin, and Stein, Benno (2012). "Paraphrase Acquisition via Crowdsourcing and Machine Learning". In: *Transactions on Intelligent Systems and Technology (ACM TIST)*.

Buzek, Olivia, Resnik, Philip, and Bederson, Benjamin B (2010). "Error driven paraphrase annotation using mechanical turk". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, pp. 217–221.

Callison-Burch, Chris (2008). "Syntactic constraints on paraphrases extracted from parallel corpora". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Celikyilmaz, A., Hakkani-Tur, D., and Feng, J. (2010). "Probabilistic model-based sentiment analysis of twitter messages". In: *Spoken Language Technology Workshop*. IEEE, pp. 79–84.

Chakrabarti, D. and Punera, K. (2011). "Event summarization using tweets". In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 66–73.

Chambers, John Kenneth and Schilling, Natalie (2013). *The handbook of language variation and change*. Vol. 80. Wiley-Blackwell.

Chandrasekar, R., Doran, Christine, and Srinivas, B. (1996). "Motivations and methods for text simplification". In: *Proceedings of the 16th conference on Computational linguistics*.

BIBLIOGRAPHY

Chen, David L. and Dolan, William B. (2011). "Collecting Highly Parallel Data for Paraphrase Evaluation". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, OR.

Chunara, Rumi, Andrews, Jason R., and Brownstein, John S. (2012). "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak". In: *The American Journal of Tropical Medicine and Hygiene* 86.1, pp. 39–45.

Clarke, J. and Lapata, M. (2008). "Global inference for sentence compression: An integer linear programming approach". In: *Journal of Artificial Intelligence Research* 31.1, pp. 399–429.

Clough, P., Gaizauskas, R., and Piao, SL (2002). "Building and annotating a corpus for the study of journalistic text reuse". In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Vol. 5, pp. 1678–1691.

Cohn, T., Callison-Burch, C., and Lapata, M. (2008). "Constructing corpora for the development and evaluation of paraphrase systems". In: *Computational Linguistics* 34.4, pp. 597–614.

Cohn, T. and Lapata, M. (2009). "Sentence compression as tree transduction". In: *Journal of Artificial Intelligence Research* 34, pp. 637–674.

Das, Dipanjan and Smith, Noah A (2009). "Paraphrase identification as probabilistic quasi-synchronous recognition". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*. Association for Computational Linguistics.

Daumé III, H. and Marcu, D. (2006). "Domain adaptation for statistical classifiers". In: *Journal of Artificial Intelligence Research* 26.1, pp. 101–126.

BIBLIOGRAPHY

Deléger, L. and Zweigenbaum, P. (2009). "Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora". In: *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*. Association for Computational Linguistics, pp. 2–10.

Denkowski, Michael, Al-Haj, Hassan, and Lavie, Alon (2010). "Turker-Assisted Paraphrasing for English-Arabic Machine Translation". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Los Angeles: Association for Computational Linguistics.

Dolan, Bill, Quirk, Chris, and Brockett, Chris (2004). "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources". In: *Proceedings of the 20th International Conference on Computational Linguistics*.

Dolan, W.B. and Brockett, C. (2005). "Automatically constructing a corpus of sentential paraphrases". In: *Proceedings of the 3rd International Workshop on Paraphrasing*.

Eisenstein, Jacob, O'Connor, Brendan, Smith, Noah A., and Xing, Eric P. (2010). "A latent variable model for geographic lexical variation". In: *Proceedings of the 2010 Conference on Empirical Methods on Natural Language Processing*.

Elhadad, N. and Sutaria, K. (2007). "Mining a lexicon of technical terms and lay equivalents". In: *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*.

Elliott, Ward EY and Valenza, Robert J (2011). "Shakespeare's Vocabulary: Did It Dwarf All Others?" In: *Stylistics and Shakespeare's Language: Transdisciplinary Approaches*, pp. 34–57.

Fader, Anthony, Zettlemoyer, Luke, and Etzioni, Oren (2013). "Paraphrase-Driven Learning for Open Question Answering". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

BIBLIOGRAPHY

Fellbaum, Christiane (2010). *WordNet*. Springer.

Foster, J., Çetinoglu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., Van Genabith, J., et al. (2011). "# hardtoparse: POS Tagging and Parsing the Twitterverse". In: *proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, pp. 20–25.

Francis, W Nelson and Kucera, Henry (1979). "Brown corpus manual". In: *Brown University Department of Linguistics*.

Fung, P. and Cheung, P. (2004a). "Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Vol. 4, pp. 25–26.

— (2004b). "Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus". In: *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, p. 1051.

Gamon, Michael (2004). "Linguistic correlates of style: authorship classification with deep linguistic analysis features". In: *Proceedings of the 20th international conference on Computational Linguistics*.

Gamon, Michael, Gao, Jianfeng, Brockett, Chris, Klementiev, Alexander, Dolan, William B., Belenko, Dmitriy, and Vanderwende, Lucy (2008). "Using contextual speller techniques and language modeling for ESL error correction". In: *Proceedings of the Third International Joint Conference on Natural Language Processing*.

Ganitkevitch, Juri, Callison-Burch, Chris, Napoles, Courtney, and Van Durme, Benjamin (2011). "Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1168–1179.

BIBLIOGRAPHY

Ganitkevitch, Juri, Van Durme, Benjamin, and Callison-Burch, Chris (2013). "PPDB: The Paraphrase Database". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 758–764.

Giampiccolo, Danilo, Magnini, Bernardo, Dagan, Ido, and Dolan, Bill (2007). "The third PASCAL recognizing textual entailment challenge". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Gimpel, Kevin, Schneider, Nathan, O'Connor, Brendan, Das, Dipanjan, Mills, Daniel, Eisenstein, Jacob, Heilman, Michael, Yogatama, Dani, Flanigan, Jeffrey, and Smith, Noah A. (2011). "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Gouws, S., Hovy, D., and Metzler, D. (2011). "Unsupervised mining of lexical variants from noisy text". In: *Proceedings of the First workshop on Unsupervised Learning in NLP*. Association for Computational Linguistics, pp. 82–90.

Greene, Erica, Bodrumlu, Tugba, and Knight, Kevin (2010). "Automatic analysis of rhythmic poetry with applications to generation and translation". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Guo, Weiwei and Diab, Mona (2012). "Modeling sentences in the latent space". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Guo, Weiwei, Li, Hao, Ji, Heng, and Diab, Mona (2013). "Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media". In: *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*.

BIBLIOGRAPHY

Han, Bo and Baldwin, Timothy (2011). "Lexical normalisation of short text messages: Makn sens a# twitter". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies*. Vol. 1, pp. 368–378.

Han, Bo, Cook, Paul, and Baldwin, T. (2012). "Automatically Constructing a Normalisation Dictionary for Microblogs". In: *Proceedings of the 2012 Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*.

Ho, ChukFong, Murad, Masrah Azrifah Azmi, Doraisamy, Shyamala, and Kadir, Rabiah Abdul (2012). "Extracting lexical and phrasal paraphrases: a review of the literature". In: *Artificial Intelligence Review*, pp. 1–44.

Hoffmann, Raphael, Zhang, Congle, Ling, Xiao, Zettlemoyer, Luke S., and Weld, Daniel S. (2011). "Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Hovy, Eduard, Marcus, Mitchell, Palmer, Martha, Ramshaw, Lance, and Weischedel, Ralph (2006). "OntoNotes: the 90% solution". In: *Proceedings of the Human Language Technology conference - North American chapter of the Association for Computational Linguistics Annual Meeting*. Association for Computational Linguistics.

Ibrahim, Ali (2002). "Extracting paraphrases from aligned corpora". PhD thesis. Massachusetts Institute of Technology.

Ibrahim, A., Katz, B., and Lin, J. (2003). "Extracting structural paraphrases from aligned monolingual corpora". In: *Proceedings of the second international workshop on Paraphrasing-Volume 16*. Association for Computational Linguistics, pp. 57–64.

Jaccard, P. (1912). "The Distribution of the Flora in the Alpine Zone". In: *New Phytologist* 11.2, pp. 37–50.

BIBLIOGRAPHY

Jiang, J. and Zhai, C.X. (2006). "Exploiting domain structure for named entity recognition". In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, pp. 74–81.

Jiang, Jing (2008). *A literature survey on domain adaptation of statistical classifiers*. URL: http://sifaka.cs.uiuc.edu/jiang4/domainadaptation/survey/da_survey.pdf.

Johnson, J.H., Martin, J., Foster, G., and Kuhn, R. (2007). "Improving translation quality by discarding most of the phrasetable". In: *Proceedings of the 2007 Joint Meeting of the Conference on Empirical Methods on Natural Language Processing*.

Joseph, Fleiss L (1971). "Measuring nominal scale agreement among many raters". In: *Psychological Bulletin* 76.5, pp. 378–382.

Kaufmann, Max and Kalita, Jugal (2010). "Syntactic normalization of Twitter messages". In: *Proceedings of the International Conference on Natural Language Processing*.

Ke, J., Gong, T., and Wang, W.S.Y. (2008). "Language change and social networks". In: *Communications in Computational Physics* 3.4, pp. 935–949.

Kellerer, Hans, Pferschy, Ulrich, and Pisinger, David (2004). *Knapsack problems*. Springer.

Knight, K. and Marcu, D. (2002). "Summarization beyond sentence extraction: A probabilistic approach to sentence compression". In: *Artificial Intelligence* 139.1, pp. 91–107.

Koehn, Philipp and Knight, Kevin (2000). "Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm". In: *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*.

BIBLIOGRAPHY

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra, and Herbst, Evan (2007). "Moses: open source toolkit for statistical machine translation". In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*.

Kok, Stanley and Brockett, Chris (2010). "Hitting the right paraphrases in good time". In: *Proceedings of the Human Language Technologies - the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Kwak, Haewoon, Lee, Changhyun, Park, Hosung, and Moon, Sue (2010). "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th International Conference on World Wide Web*. ACM, pp. 591–600.

Landis, J Richard and Koch, Gary G (1977). "The measurement of observer agreement for categorical data". In: *Biometrics*, pp. 159–174.

Lee, L., Aw, A., Zhang, M., and Li, H. (2010). "Em-based hybrid model for bilingual terminology extraction from comparable corpora". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pp. 639–646.

Li, Wenjie, Wu, Mingli, Lu, Qin, Xu, Wei, and Yuan, Chunfa (2006). "Extractive summarization using inter- and intra- event relevance". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, pp. 369–376.

Li, Linlin and Sporleder, Caroline (2009). "Classifier combination for contextual idiom detection without labelled data". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 315–323.

BIBLIOGRAPHY

Li, Qi (2012). *Literature survey: Domain Adaptation Algorithms for Natural Language Processing*. URL: http://nlp.cs.rpi.edu/paper/qisurvey.pdf?.

Lin, D. and Pantel, P. (2001). "Discovery of inference rules for question-answering". In: *Natural Language Engineering* 7.4, pp. 343–360.

Liu, X., Li, K., Han, B., Zhou, M., Jiang, L., Xiong, Z., and Huang, C. (2010). "Semantic role labeling for news tweets". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 698–706.

Liu, X., Li, K., Zhou, M., and Xiong, Z. (2011c). "Enhancing semantic role labeling for tweets using self-training". In: *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence*.

Liu, F., Weng, F., Wang, B., and Liu, Y. (2011a). "Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision". In: *The 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies*.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011d). "Recognizing named entities in tweets". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies*, pp. 359–367.

Liu, F., Liu, Y., and Weng, F. (2011b). "Why is 'sxsw' trending? exploring multiple text sources for twitter topic summarization". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies*, p. 66.

Liu, Fei, Weng, Fuliang, and Jiang, Xiao (2012). "A broadcoverage normalization system for social media language". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

BIBLIOGRAPHY

Madnani, Nitin and Dorr, Bonnie J. (2010). "Generating phrasal and sentential paraphrases: A survey of data-driven methods". In: *Computational Linguistics* 36.3, pp. 341–387.

Mani, Inderjeet and Wilson, George (2000). "Robust temporal processing of news". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, pp. 69–76.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). "Domain adaptation: Learning bounds and algorithms". In: *arXiv preprint arXiv:0902.3430*.

McKeown, K.R. (1979). "Paraphrasing using given and new information in a question-answer system". In: *Proceedings of the 17th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 67–72.

Mehta, Cyrus R and Patel, Nitin R (1997). "Exact inference for categorical data". In: *Biometrics* 53.1, pp. 112–117.

Och, Franz Josef (2003). "Minimum error rate training in statistical machine translation". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*.

Mintz, Mike, Bills, Steven, Snow, Rion, and Jurafsky, Dan (2009). "Distant supervision for relation extraction without labeled data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pp. 1003–1011.

Moore, Robert C. (2002). "Fast and Accurate Sentence Alignment of Bilingual Corpora". In: *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*.

Moore, Robert C. (2004). "On log-likelihood-ratios and the significance of rare events". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 333–340.

Mota, C. and Grishman, R. (2009). "Updating a name tagger using contemporary unlabeled data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Association for Computational Linguistics, pp. 353–356.

— (2008). "Is this NE tagger getting old?" In: *Proceedings of the International Conference on Language Resources and Evaluation*.

Muzny, Grace and Zettlemoyer, Luke (2013). "Automatic Idiom Identification in Wiktionary". In: *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

Naaman, M., Becker, H., and Gravano, L. (2011). "Hip and trendy: Characterizing emerging trends on Twitter". In: *Journal of the American Society for Information Science and Technology* 62.5, pp. 902–918.

Nelken, R. and Shieber, S.M. (2006). "Towards robust context-sensitive sentence alignment for monolingual corpora". In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 161–168.

Nenkova, Ani and Vanderwende, Lucy (2005). *The impact of frequency on summarization*. Tech. rep. MSR-TR-2005-101. Redmond, Washington: Microsoft Research.

Och, Franz Josef and Ney, Hermann (2003). "A systematic comparison of various statistical alignment models". In: *Computational linguistics* 29.1, pp. 19–51.

BIBLIOGRAPHY

Paşca, M. and Dienes, P. (2005). "Aligning needles in a haystack: Paraphrase acquisition across the web". In: *Proceedings of the Second international joint conference on Natural Language Processing*, pp. 119–130.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Petrović, Saša, Osborne, Miles, and Lavrenko, Victor (2012). "Using paraphrases for improving first story detection in news and Twitter". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*.

Quirk, Chris, Brockett, Chris, and Dolan, William (2004). "Monolingual Machine Translation for Paraphrase Generation". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Raghavan, Sindhu, Kovashka, Adriana, and Mooney, Raymond (2010). "Authorship attribution using probabilistic context-free grammars". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Riedel, Sebastian, Yao, Limin, and McCallum, Andrew (2010). "Modeling Relations and Their Mentions without Labeled Text". In: *Proceedigns of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 148–163.

Ritter, Alan, Cherry, Colin, and Dolan, William B. (2011a). "Data-driven response generation in social media". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ritter, Alan, Clark, Sam, Mausam, and Etzioni, Oren (2011b). "Named entity recognition in tweets: an experimental study". In: *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1524–1534.

Ritter, Alan, Mausam, Etzioni, Oren, and Clark, Sam (2012). "Open domain event extraction from twitter". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1104–1112.

Ritter, Alan, Zettlemoyer, Luke, Mausam, and Etzioni, Oren (2013). "Modeling Missing Data in Distant Supervision for Information Extraction". In: *Transactions of the Association for Computational Linguistics* 1, pp. 367–378.

Robbins, Herbert (1985). "Some aspects of the sequential design of experiments". In: *Herbert Robbins Selected Papers*. Springer, pp. 169–177.

Rumšienė, Goda (2004). "Development of Internet English: alternative lexis, syntax and morphology". In: *Kalbu Studijos* 6, p. 48.

Wu, Dekai, Addanki, Karteek, Saers, Markus, and Beloucif, Meriem (2013). "Learning to Freestyle: Hip Hop Challenge-Response Induction via Transduction Rule Segmentation". In: *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

Sekine, S. (2005). "Automatic paraphrase discovery based on context and keywords between ne pairs". In: *Proceedings of the 3rd International Workshop on Paraphrasing*. Vol. 2005. Jeju Island, pp. 4–6.

Shima, Hideki and Mitamura, Teruko (2011). "Diversity-aware evaluation for paraphrase patterns". In: *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*. Association for Computational Linguistics, pp. 35–39.

Shinyama, Yusuke and Sekine, Satoshi (2003). "Paraphrase acquisition for information extraction". In: *Proceedings of the second international workshop on Paraphrasing*.

BIBLIOGRAPHY

Shinyama, Y., Sekine, S., and Sudo, K. (2002). "Automatic paraphrase acquisition from news articles". In: *Proceedings of the second international conference on Human Language Technology Research*.

Smith, J.R., Quirk, C., and Toutanova, K. (2010). "Extracting parallel sentences from comparable corpora using document level alignment". In: *Proceedings of the Human Language Technologies - The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 403–411.

Snow, Rion, O'Connor, Brendan, Jurafsky, Daniel, and Ng, Andrew Y. (2008). "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Spreyer, Kathrin and Kuhn, Jonas (2009). "Data-driven dependency parsing of new languages using incomplete and noisy training data". In: *Proceedings of the 13th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 12–20.

Stolcke, Andreas (2002). "SRILM - an extensible language modeling toolkit". In: *Proceedings of the International Conference on Spoken Language Processing*.

Subramaniam, L.V., Roy, S., Faruquie, T.A., and Negi, S. (2009). "A survey of types of text noise and techniques to handle noisy text". In: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. ACM, pp. 115–122.

Surdeanu, Mihai, Tibshirani, Julie, Nallapati, Ramesh, and Manning, Christopher D (2012). "Multi-instance multi-label learning for relation extraction". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 455–465.

Tran-Thanh, Long, Stein, Sebastian, Rogers, Alex, and Jennings, Nicholas R (2012). "Efficient crowdsourcing of unknown experts using multi-armed bandits". In: *European Conference on Artificial Intelligence*, pp. 768–773.

Vanderwende, Lucy, Suzuki, Hisami, Brockett, Chris, and Nenkova, Ani (2007). "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion". In: *Information Processing & Management* 43.6, pp. 1606–1618.

Vermorel, Joannes and Mohri, Mehryar (2005). "Multi-armed bandit algorithms and empirical evaluation". In: *Machine Learning: ECML 2005*. Springer, pp. 437–448.

Volkova, Svitlana, Wilson, Theresa, and Yarowsky, David (2013). "Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media". In: *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

Wan, Stephen, Dras, Mark, Dale, Robert, and Paris, Cécile (2006). "Using dependency-based features to take the "para-farc" out of paraphrase". In: *Proceedings of the Australasian Language Technology Workshop*. Vol. 2006.

Xu, Wei, Grishman, Ralph, and Zhao, Le (2011b). "Passage retrieval for information extraction using distant supervision". In: *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 1046–1054.

Xu, Wei and Grishman, Ralph (2009). "A parse-and-trim approach with information significance for Chinese sentence compression". In: *Proceedings of the 2009 Workshop on Language Generation and Summarisation*. Association for Computational Linguistics, pp. 48–55.

Xu, Wei, Tetreault, Joel, Chodorow, Martin, Grishman, Ralph, and Zhao, Le (2011a). "Exploiting syntactic and distributional information for spelling correction with web-

scale N-gram models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1291–1300.

Xu, Jun-Ming, Jun, Kwang-Sung, Zhu, Xiaojin, and Bellmore, Amy (2012a). "Learning from bullying traces in social media". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*.

Xu, Wei, Ritter, Alan, Dolan, Bill, Grishman, Ralph, and Colin, Cherry (2012b). "Paraphrasing for style". In: *Proceedings of the 24th International Conference on Computational Linguistics*.

Xu, Wei, Ritter, Alan, and Grishman, Ralph (2013c). "Gathering and generating paraphrases from twitter with application to normalization". In: *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, pp. 121–128.

Xu, Wei, Grishman, Ralph, Meyers, Adam, and Ritter, Alan (2013a). "A Preliminary Study of Tweet Summarization using Information Extraction". In: *Proceedings of the NAACL Workshop on Language Analysis in Social Media*.

Xu, Wei, Hoffmann, Raphael, Zhao, Le, and Grishman, Ralph (2013b). "Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Yao, Xuchen, Van Durme, Benjamin, and Clark, Peter (2013). "Semi-Markov Phrase-based Monolingual Alignment". In: *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). "Inducing multilingual text analysis tools via robust projection across aligned corpora". In: *Proceedings of the first*

international conference on Human language technology research. Association for Computational Linguistics, pp. 1–8.

Yatskar, Mark, Pang, Bo, Danescu-Niculescu-Mizil, Cristian, and Lee, Lillian (2010). "For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia". In: *Proceedings of the Human Language Technologies - The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Zanzotto, Fabio Massimo, Pennacchiotti, Marco, and Tsioutsiouliklis, Kostas (2011). "Linguistic redundancy in twitter". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 659–669.

Zesch, Torsten, Müller, Christof, and Gurevych, Iryna (2008). "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary." In: *Proceedings of the International Conference on Language Resources and Evaluation*. Vol. 8, pp. 1646–1652.

Zhao, S., Wang, H., Liu, T., and Li, S. (2008). "Pivot approach for extracting paraphrase patterns from bilingual corpora". In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies*, pp. 780–788.