Evaluating Open Relation Extraction Over Conversational Texts

by

Mahsa Imani

B.Sc., Tarbiat Moalem University of Tehran, 2008
M.Sc., Isfahan University of Technology, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

January 2014

© Mahsa Imani 2014

Abstract

In this thesis, for the first time the performance of Open IE systems on conversational data has been studied. Due to lack of test datasets in this domain, a method for creating the test dataset covering a wide range of conversational data has been proposed. Conversational text is more complex and challenging for relation extraction because of its cryptic content and ungrammatical colloquial language. As a consequence text simplification has been used as a remedy to empower Open IE tools for relation extraction. Experimental results show that text simplification helps OLLIE, a state of the art for relation extraction, find new relations, extract more accurate relations and assign higher confidence scores to correct relations and lower confidence scores to incorrect relations for most datasets. Results also show some conversational modalities such as emails and blogs are easier for relation extraction task while people reviews on products is the most difficult modality.

Preface

This dissertation is original, unpublished, independent work by the author, Mahsa Imani. I designed the research project with the help of Giuseppe Carenini and Yashar Mehdad. I proposed an approach for mitigating the problems faced in this research area. I conducted all the experiments and evaluated the performance of the proposed approach. I analyzed the experimental data and wrote the whole thesis. Giuseppe Carenini and Yashar Mehdad were the supervisory authors on this project and were involved throughout the project in concept formation and manuscript edits.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my special appreciation and thanks to my advisor, Dr. Giuseppe Carenini, who introduced me to Natural Language Processing and allowed me to grow as a research scientist in this area. His advice, knowledge and encouragement allowed me to develop and pursue this thesis.

I would also like to express my gratitude to Dr. Reymanod Ng and Dr. Yashar Mehdad for their valuable advice, help, and guidance.

Last but not least, I would like to thank my friends and family for their love and support during all my studies.

Dedication

To my mother, Ehteram Jafari, for her unconditional love and support throughout my life.

Chapter 1

Introduction

1.1 Motivation

In past human could only interact and communicate with the people they know by speaking or writing letters about events, concepts and ideas. With the invention of internet and prevalence of email systems, blogs, fora discussions and social networking, now people who even dont know each other can participate in different conversations and discuss their thoughts, feelings and opinions. They can ask any question in social streams or online discussion groups and find the answer by reading and analyzing the comments posted by different people around the world. They can discuss new products and services and make an informed decision.

The conversational data is growing in an exponential rate. Everyday new reviews are written and new discussions are ongoing in social media about products, services and events and nobody is able to read them all and make an informative summary of them. People may want to join a discussion held by more than 100 people and need to know what details have been discussed by the time they joined. To take advantage of this massive conversational data, we need new tools to help us summarize and find relevant information.

To help people find what is closer to their information need, we need new tools to deal with this data explosion. To effectively manage, summarize, search and find relevant information, structured knowledge is required. Relation extraction is the task of finding relationships between entities in text and is an effective way to convert unstructured text data such as blogs, web pages, news, scientific literature, and online reviews into structured knowledge. This structured knowledge offers users and organizations an effective and novel way to get and analyze the information they need to achieve their goals.

There are many other scenarios in which we are interested in discovering relationships within a set of entities in documents. Relations can be used for finding gene-disease relationships [16], finding the relationships between drugs, genes/proteins and diseases [28], question answering [48, 58], summarization [51], automatic database generation, intelligent document searching,

ranking and indexing [3], ontology population [39, 43, 55, 58], and finding protein-protein interactions [29, 33].

1.2 Open Information Extraction and Its Challenges

In Traditional Information Extraction (IE), the relation of interest has to be specified in advance. One has to provide those systems with new extraction patterns or training examples for new relations. These systems require one pass over corpus for each relation and hence, they are not scalable with the size and variety of the web corpus [5]. Open IE systems address this problem by extracting relations from arbitrary sentences without requiring domain-specific knowledge and target relations in advance [5].

Open IE systems are scalable in a sense that they extract various relations in a single pass or few numbers of passes over the corpus [5]. State of-the-art Open IE systems such as ReVerb [24], WOE [62], OLLIE [38], SONEX [42], TreeKernel [63], EXEMPLAR [19] extract web-scale information in the format of relational tuples (*arg1; rel; arg2*) in which the relation phrase *rel* expresses a relation between *arg1* and *arg2*.

There are many challenges in extracting semantic relationships between entities. The most important one is the variety of relation forms which makes it very difficult to be effictively learnt through machine learning approaches or to be captured through regular expressions and rule based systems: Relations can be synonymous [5, 48, 64], negated [40], n-ary [40], conditionally true [38], infrequent [21], or implicit [63]; They can have light verb constructs [24], non-contiguous relation phrases [24], or subsume other relations [58]. Not only relation forms are various but also their arguments. Arguments can also be synonymous and have different forms. They can be a Noun Phrase (NP) [24, 38, 62], a Named Entity (NE) [19, 42], or even a sentence[1].

All these challenges have been addressed in the literature but each relation extraction approach tackles a subset of these problems. Long, compound and complex sentences pose new challenges and make current extraction approaches considerably less effective. These tools fail to find all the relations when a sentence contains relative clause modifier, referent, or relative relations [24].

Open IE can be utilized to convert the massive amount of available conversational data into structured knowledge and as a consequence help

[1] [15] introduced nested relations in which one of the arguments is a sentence.

us summarize, search and find relevant information. But conversational text poses new challenges for Open IE due to its specific characteristics including cryptic content, lots of abbreviations, ungrammatical and informal language. The problems arise from difficulty in parsing the sentence at the preprocessing step to extracting relations themselves.

1.3 Conversational Data and New Challenges

Conversational data is growing in an exponential rate in the forms of Emails, blogs, reviews, meeting records, or posts in social streams [9]. This data is an invaluable source of information. It provides organizations and people with public feelings and opinions towards new products, services, and events [45]. As a consequence, there is an ongoing research on conversational data in order to represent the content of these conversations in an informative way to summarize them, find the relevant information and the content worth reading e.g. [35, 41, 44, 57].

Sentences in conversational data such as social streams, chat logs, blogs and Email threads are complex and noise-prone [9]. They often have an ungrammatical colloquial language, more abbreviations, and may not state the full relation which is often assumed in relation extraction task. Hence they pose new challenges and make current extraction approaches considerably less effective. There is another challenge in applying these techniques designed for extracting relations from non-conversational well-written text to conversational text. Performance of these techniques depends on the output of preprocessing steps such as Part Of Speech (POS) tagging, NP chunking, NE tagging and dependency parsing whose accuracy degrade for conversational text. About 8% of missed extractions and 7-32% of incorrect extractions in ReVerb, WOE-parse and OLIIE are due to incorrect parsing [24, 62]. Apart from these challenges, sentences can be simplified by following a set of lexical and syntactic rules [14, 18, 52] or log-linear models [4]. Text simplification can improve the accuracy of preprocessing steps [13, 14, 52] as well as relation extraction by breaking down each complex sentence into semantically equivalent shorter sentences.

1.4 Text Simplification

Text simplification is the process of simplifying texts while preserving their meaning and information to increase understandability or make it easier to process by computers [4]. Text simplification can be syntactic or lexical. To

simplify texts lexically difficult words are substituted by easier words. For syntactic simplification, a set of rules [14, 18, 52] or log-linear models [4] can be utilized to simplify sentences by breaking down them into shorter and simpler sentences.

Text simplification has been studied as a preprocessing step for several Natural Language Processing (NLP) tasks such as relation extraction [33], semantic role labelling (SRL) [61], machine translation [46], summarization [53], and improving the accuracy of parsers [13, 52]. Preprocessing text to simplify it has been inspired by the fact that performance of these systems rapidly deteriorates as the length and complexity of the sentence increases [13, 33].

Most of the errors in parsing are due to long, complex, and ambiguous sentences and it has been shown that text simplification and compression eases summarization by converting complex long sentences into shorter sentences and dropping non-essential information [53]. Performance of Open IE systems depends on the output of preprocessing step such as POS tagging, NP chunking, NE tagging and dependency parsing whose accuracy degrade for complex conversational text. Syntactically simplifying texts will lead to more accurate sentence level analysis and hence a more accurate relation extraction.

Jonnalagadda and Gonzalez [33] showed that sentence simplification considerably helps relation extraction in the domain of biomedical texts which usually have longer sentences with more abbreviations and relative clauses than less specialized and less technical texts like news. As opposed to scientific literature in which sentences are grammatically correct, sentences in conversational texts are not well-written. They are noise-prone and contain ungrammatical text with much cryptic content. But in both domains, more abbreviations than general text are used. As a consequence, we hypostatize that text simplification may be of benefit in the domain of conversational data as well. It is much more challenging to extract correct relations from compound, long and syntactically ambiguous sentences. By breaking down sentences into shorter and simpler sentences, we empower relation extraction tools to extract more relations and the extracted relations will be more accurate.

1.5 Problem Statement and Contribution

The purpose of this study is to investigate the performance of open relation extraction tools on conversational data for the first time and suggest meth-

ods to tackle the challenges faced in this domain. In particular, the effect of text simplification before relation extraction will be evaluated in the domain of conversational texts. For this purpose, first a test dataset covering a wide range of conversational data and sentences has been populated from different corpora. The dataset creation approach has been described in details in chapter 3. Then the performance of Ollie, a state of the art for relation extraction, will be evaluated on the test dataset sampled from Emails, tweets, product reviews and blogs corpora before and after simplification based on the number of extracted relations, the accuracy of extracted arguments and relation phrases, and confidence score of extracted relations. We refer to the later system (OLLIE using text simplification as a preprocessing step) as OLLIE-Simplified. For simplification TriS has been utilized to syntactically simplify sentences before relation extraction [4]. There are other Open IE tools such as TreeKernel, SONEX and EXEMPLAR with better reported accuracy than OLIIE. We were not able to use them for our datasets due to the fact that they extract relations between named entities and were able to extract less than five relations from each dataset while OLIIE extract hundreds of relations between noun phrases. In addition, TreeKernel limits the domain of its usage because of its supervised approach.

We show text simplification is of great benefit in empowering relation extraction in the domain of conversational data. Experimental results show that after text simplification by TriS, OLLIE-Simplified outperforms OLLIE in terms of accuracy and informativeness of confidence score. It assigns higher confidence scores to correct relations and in most cases lower confidence scores to incorrect relations. Experimental results also suggest that a new system which utilizes the union of extracted relations of two systems will outperform both systems, OLLIE and OLLIE-Simplified, in terms of recall since each system can find distinct relations not found by the other one.

In summary the three main contributions are as follows:

- Collecting and sampling a dataset covering different conversation modalities.
- For the first time evaluating the performance of Open IE in the area of conversational texts over the created dataset.
- Evaluating the performance of Open IE on conversations after text simplification.

1.6 Outline

The outline of the thesis is as follows: In the next chapter, available conversational datasets, relation extraction and text simplification approaches are reviewed. Chapter 3 describes our methodology for creating a test dataset and Open IE. In chapter 4, experiments are described and two systems are compared. At the end in chapter 5, conclusion and future works are presented.

Chapter 2

Background and Related Work

In this chapter, first available conversational datasets have been reviewed. Then relation extraction approaches (traditional and open) have been briefly reviewed. Then text simplification which has been used as a preprocessing step has been described at the end.

2.1 Conversational Datasets

There are different conversation modalities or domains including chats, emails, meetings and blogs which are distinguishable by different characteristics they have. Conversations can be categorized into two groups of synchronous and asynchronous. In synchronous conversations such as meetings and chats, turns happen with minimal gap and overlap between them. In asynchronous conversations such as fora discussions, emails, microblogs and blogs, different people can participate at different times or even same time making a more complicated conversational structure [9].

The length of turns in different conversation modalities vary. While there is no limit on the length of turns in synchronous conversations, length of other modalities are usually limited. For example, in twitter, each tweet must be 140 character long and as a consequence tweets are much more cryptic and concise with more abbreviations than other modalities.

Due to extensive research on conversational data for summarization and opinion mining, there are several publicly available datasets for most modalities but there is no dataset covering all different types of conversation domains. Available corpora are as follows:

- Meetings: AMI and ICSI corpora. AMI corpus consists of 100 hours of scenario and non-scenario meetings [10]. ICSI corpus consists of 75 non-scenario or natural technical meetings held by ICSI researchers [30].

- Chats: Tux4kidss chat logs. Tux4Kids develops free software for educational purposes. The target users are kids. The dataset consists of four chat threads in plain text format. In these chat sessions free-software and educational topics as well as Tux4Kids business are discussed.[2]

- Social networks and microblogs: Because of privacy concerns, microblogs such as tweets and people's posts on other social networks such as Facebook are not publicly available or there is a limit on the number of posts which can be downloaded using their API.

- Emails: W3C, BC3, and Enron corpora. BC3 Email dataset was originally developed for summarization and contains 40 email threads and 261 Emails from W3C corpus [59]. Enron corpus contains natural emails written by 150 employees in Enron corporation [37].

- Reviews: Among datasets for reviews on products and services is Opinosis Dataset which was originally developed for summarization and contains reviews on 51 topics. Others are customer reviews on 5 products [36], amazon product reviews [31], and movie review dataset released by Pang and Lee[3].

- Blogs: There are several blog datasets including Spinn3r blog dataset[4] and Splog Blog Dataset[5].

2.2 Relation Extraction

2.2.1 Introduction

Relation extraction is the task of finding semantic relationships between a set of entities in text. Relation extraction approaches can be divided into two categories: Traditional IE and Open IE. In Traditional IE, the relation of interest has to be specified in advance while in Open IE, various relations can be extracted without requiring any prior knowledge. In the next two sections both approaches have been described.

[2] http://www.geekcomix.com/tux4kids/chatlogs/
[3] https://bitbucket.org/speriosu/updown/wiki/Corpora
[4] http://snap.stanford.edu/data/other.html
[5] http://ebiquity.umbc.edu/resource/html/id/212/Splog-Blog-Dataset

2.2.2 Traditional IE

In Traditional IE systems the relation of interest has to be specified in advance. To extract a specific relation, some of these systems use hand-coded extraction patterns or semi-supervised (bootsrapping) machine learning approaches to learn extraction patterns using a few seed instances of the relation. Among those systems are DIPRE [7], Snowball [2], KnowItAll [22], Espresso [47], Leila [54], SRES [50], Luchs [27]. Supervised relation-specific approaches look at relation extraction as a binary classification task to identify whether there is a relation between two entities or not e.g. [26, 40, 66]. One has to provide those systems with new extraction patterns and training examples for new relations. They also require one pass over corpus for each relation and hence, these approaches are not scalable with the size and variety of the web corpus [5].

DIRPE [7] finds all occurrences of seed pairs, which represent the arguments of a given relation, and construct a 6-tuple for each occurrence in the corpus. Then it induces new patterns of the relation by grouping those tuples based on the order and context between the arguments. In the next iteration, the new patterns are used to find new seed pairs and hence new patterns for the relation. This procedure will continue till some criteria are met.

Snowball [2] is an improvement over DIRPE with the difference that each pattern is represented by a 5-tuple including context before, between, and after a pair of named entity tags. In a single pass they cluster the found tuples and create a controid pattern for each cluster to be used in the next iteration for finding new instances of the relation. After each iteration, they evaluate each pattern by its precision in extracting relation instances. In the same way, each new instance in the next iteration is evaluated based on the pattern used to extract the relation and its similarity to the pattern according to the similarity function. Unlike DIRPE, Snowball evaluates and filters new patterns which helps it prevent noise propagation. It has a more flexible pattern matching approach compared to DIPRE but suffers from too many specific patterns it generates because of the way they represent a pattern.

The main goal of KnowItAll [22] is to extract entities (unary predicates). It uses generic patterns to learn domain-specific extraction rules. To extract entities, It accepts a set of entity classes like *city* and outputs entity instances extracted from web. It uses the extraction frequency as a mean to evaluate the likelihood of the extraction.

Mcdonald et al. [40] take a supervised approach and first build a feature

vector for each two named entities based on shallow syntactic structure. They find binary relations by classifying the feature vectors as related or not. Then they solve the problem of complex relations (n-ary relations) by constructing a weighted graph in which nodes are entities and the weight of edges show the confidence of having a binary relation between those entities. Then they convert any maximal clique in the graph with geometrical mean greater than 0.5 into a n-ary relation.

2.2.3 Open IE

Open IE systems address problems we face when using Traditional IE systems by extracting relations from arbitrary sentences without requiring domain-specific knowledge and target relations in advance. Open IE systems extract web-scale information in the format of relational tuples (*arg1; rel; arg2*) in one pass or constant number of passes over the corpus. The relation phrase *rel* expresses a relation between two arguments *arg1* and *arg2*.

Common relation extraction subtasks in Open IE are as follows:

1. Preprocessing steps for sentence level analysis including chunking and parsing [1]
2. Identifying arguments: usually noun phrases [62] or named entities in the sentence are considered as potential candidates [19, 42].
3. Identifying relation phrase (predicate): through learnt rules, hand-coded extraction patterns [22], machine learning, or hybrid [1].
4. Postprocessing and integrating information: including argument and relation resolution [64], co-reference resolution, deduplication, disambiguation [1, 55].
5. Identifying confidence of the extracted relations: pointwise mutual information (PMI) [22], noisy-or model, urns or contextual similarity [21].

Open IE systems differ in the order of step 2 and 3. ReVerb [24] and its extension R2A2 [23], opposed to other systems, first extracts relation phrase and then its arguments.

Most Open IE tools such as TreeKernel, SONEX and EXEMPLAR dont perform the last step which is computing a confidence score for the extracted relation.

TextRunner, the first Open IE system, starts from candidate arguments, each pair of base NPs satisfying a set of constraints, and uses a binary classifier to decide whether a relation between them should be extracted. Open IE systems aim to extract all different types of relations. Since nobody can determine how many different relations exist in the world, it is not possible to provide these systems with training examples covering all the cases. As a consequence, some of these systems take a self-supervised approach[6] to generate their training data heuristically [5, 62]. To train its classifier, TextRunner uses a self supervised learner which generates positive and negative examples by using a set of heuristic constraints. It labels extractions negative if they violate any of the constraints, otherwise positive. Then they train a CRF to extract relation phrases. TextRunner merges normalized relations and counts number of occurrences in order to later assign a confidence score to each [6].

WOE [62] also starts by first identifying arguments, NPs in the sentence satisfying a set of constraints. Then it tries to extract a relation between them. It heuristically produces training data for its extractor by matching infoboxes (attribute-value pairs) and sentences in Wikipedia articles. They compare two different extractors: 1) WOE-POS like TextRunner uses trivial features like POS tags and a trained CRF to extract relation phrases. 2) WOE-Parse outputs a relation based on the shortest dependency path between two NPs [62]. Even though infoboxes are incomplete and error prone, WOE outperformed TextRunner with much higher F-measure but the runtime of WOE-Parse was 30 times more than TextRunner which may not be suitable for Open IE.

The disadvantage of approaches which first identify arguments in the relation is that they are prone to mistakenly consider a noun as an argument while it is part of the relation phrase. It is usually the case for multi word relation phrases such as "make a deal with", "has a PhD", "is a city in" etc. [24]. Inspired by this, ReVerb [24] starts from verbs in the sentence. The longest word sequences starting with these verbs which satisfy both syntactic and lexical constraints will be outputted as relation phrases. The syntactic constraints are a set of regular expressions based on POS tags and the lexical constraint simply counts the number of distinct arguments that the extracted relation takes in the corpus of 500M Web sentences. Arguments are nearest NPs in the left and right hand side of the relation phrase satisfying a set of conditions. Later they analyzed 250 random web pages and noticed that only 65% of Arg1s and 60% of Arg2s are simple

[6]Self supervised learning algorithms heuristically label their own training data.

NPs and there are a handful of other categories covering 90% of other cases. Inspired by this observation, they trained three classifiers to determine the left and right bound of Arg1 and the right bound of Arg2. They used a set of flat features like length of sentence, the context around the argument and features inspired by their analysis denoting other categories than simple NPs [23].

There are other approaches that extract relations between named entities as opposed to noun phrases and hence are less prone to mistakenly consider a noun as an argument while it is part of the relation phrase. Named entity extraction is the process of categorizing entities into predefined classes such as persons and organizations. Despite decades of research in this area, it is still far way from complete [49, 65]. There are some drawbacks in using named entities as arguments of relations. First, no named entity extraction system performs well in all domains and lots of effort is needed to get them work on other domains than the one they designed for [65]. Second, the number of named entities extracted by these systems has been restricted by the number of categories and subcategories of named entities defined. Usually there is a need to extend the range of named entity categories for a new domain [65]. Among relation extraction tools falling in this category are, SONEX, TreeKernel, and EXEMPLAR.

SONEX [42] groups sentences having the same pair of named entities and presents each such group with a vector of shallow features including unigrams, bigrams and part of speech patterns of the context between the pair of entities. Then SONEX clusters these feature vectors and assigns a label based on these features to each cluster which represents the relation phrase between the two named entities. They evaluate the extracted relations by their system by that of Freebase[7]. Working on blogosphere, their ultimate goal is to build a scalable system that outputs a social network of the entities by considering the named entities as nodes and the relation label as edges of the network.

One of the main problems in Open IE is the variety of relation forms which makes it very difficult to be effictively learnt through machine learning approaches or be captured through regular expressions and rule based systems. In addition, there are implicit relations such as "located in" between Nishapur and Iran in "Omar Khayyam was born in Nishapur, Iran" which make extractions much more challenging. To mitigate this problem, TreeKernel [63] breaks down the relation extraction task into two subtasks. In the first subtask, they extract entities and feed a SVM model with the de-

[7] `http://www.freebase.com`

pendency path between them to decide whether there is a relation between the entities. As a consequence implicit relations and all relation forms will be considered. For the second subtask, they employ regular expressions patterns based on those of ReVerb to extract several candidates for relation phrase and then utilize another SVM dependency kernel to decide whether these candidates are correct. Even though the first SVM model does not put any constraint on the relation form, the input of second one is restricted to nominal and verbal candidates. Though their approach outperformed OLLIE and ReVerb for both subtasks, lack of training examples in other domains makes their method less practical.

[15] showed how semantic role labelers can be used for open relation extraction. They convert output of these systems to equivalent relational tuples of TextRunner. According to them, TextRunner is much faster due to shallow analysis and more practical in the case of limited time but the semantic role labelers outperformed TextRunner given unlimited time. They also propose a system which makes use of the union of output of these systems and is the best given intermediate amount of time.

Mesquita et al. [19] classify relation extraction approaches into three categories based on the depth of analysis they do: The first category encompasses shallow approaches such as that of TextRunner, ReVerb, and SONEX which extract relations based on POS tags of the sentences. The second category takes advantage of dependency parse tree of the sentence; OLLIE and TreeKernel fall in this category. In their classification semantic role labelers such as Lund [32] and SwiRL [56] are another category of relation extraction approaches which do a more sophisticated analysis than dependency parsing. They discuss how increasing the complexity of analysis increases the computational cost but does not essentially lead to a high increase in the accuracy. As a consequence, they proposed EXEMPLAR, a rule based system, which utilizes the idea behind the success of semantic role labelers, considering the connection between relation words and arguments. By using dependency parse of the sentence instead of semantic role, they keep computational cost the same as the second category. They show their approach is superior to other methods when extracting relations between named-entities not noun phrases.

There are other approaches with different frameworks. One of them is an unsupervised method to semantically parse or represent meaning of a sentence, which subsumes relation extraction, proposed by Poon and Domingos [48]. In their setting, the semantic parse of the sentence is the set of fragments obtained by partitioning its syntactic dependency tree and later assigning each fragment to a cluster of semantically identical structures. Each

cluster contains structures syntactically or lexically different but having the same meaning which resolves the problem of argument resolution and relation resolution. Their goal is to find this set of clusters which corresponds to target predicates (relations) and objects (arguments). USP takes advantage of Markov Logic Networks to model the joint distribution for dependency tree and its latent meaning representation (MR). It tries to maximize the probability of the observing dependency structures of the sentence by tuning the weights of first-order clauses. In OntoUSP [58], authors modified the cluster mixture formula in USP to also include ISA relationships between clusters which leads to better generalization. For example in their setting, there is ISA relationship between inhibit and regulate.

Another interesting approach for relation extraction is the approach taken in SOFIE [55]. SOFIE first converts everything (ontology, text, constraints, and new fact hypotheses) into logical statements. Then they use logical rules to determine which hypotheses are probably true. They manually developed a set of general rules, conceptually similar to DIRPE and Snowball, but relation-specific rules can be added later. They use the weighted MAX-SAT setting to find out set of hypotheses that should be true in order to have the maximum number of rules satisfied. Their emphasis is more on ontology population and reasoning.

The problem with systems such as ReVerb, SwiRL and WOE is that there are only capable of extracting verb based relations. Verb based relations are those relations which begin with a verb. ReVerb further limits these relations to be between arguments and satisfy syntactic constraints. Even though WOE can find relations not between arguments, it fails to find those relations which contains nouns such as is CEO of. There are other types of relations beginning with other syntactic types e.g. author of and such as which they are not capable of extracting. Another weakness of these systems is that they ignore context leading to extraction of relations which are conditionally or supposedly true [38]. For example in the sentence:

If John had a million dollars, he would buy a house.

The relation (John; buy; a house) is only correct when he has a million dollars.

OLLIE tries to address the weaknesses of previous approaches by taking advantage of high confidence extractions of ReVerb as input seeds of its bootsrapper. Unlike other proposed bootsrapped methods for which seeds are a pair of arguments and bootstrapper considers those sentences that match both arguments, bootsrapper in OLLIE not only matches arguments but also relation words. It empowers bootsrapper to learn general open

extraction patterns based on dependency parse tree of retrieved sentences that can be utilized to extract other relations.

Bootsrapper of WOE-Parse also takes the same approach to learn extraction patterns. The thing that makes a big difference in their performance is the quality of their seeds. WOE-parse retrieves those sentences in Wikipedia article that matches infobox values (candidate arguments) and heuristically considers all the words between arguments as a relation phrase which does not hold true in many cases causing noisy seeds. OLLIE also introduces a context analysis component that utilizes the dependency parse tree and two simple rules to find relations which are conditionally or supposedly true and adds a filed indicating the conditional truth or attribution. If there is clausal compliment (ccomp) edge in the tree, they see whether the verb of the tree exists in a list of communication and cognition verbs. If so, they add the attribution field to the extracted relation. In the same manner, they add a causal modifier field, if there is adverbial clause (advcl) edge in the tree and the first word of clause exists in a list of 16 terms: if, when, although,. Since these two rules dont cover all the possible conditional or hypothetical true relations, they train a classifier to decrease the confidence of the relation in other cases (Mausam et al., 2011).

2.3 Text simplification

In this section, first text simplification (syntactic and lexical) approaches have been briefly reviewed. TriS, which has been used in the experiments, has been described in more details at the end.

2.3.1 Introduction

The same meaning can be expressed in many different ways with different level of complexity to understand. The source of this complexity can be syntactic or lexical. Syntactic complexity arises from complex, compound and nested structures usually in long sentences and lexical complexity arises from use of difficult and less frequent words or ambiguity in their meaning. Text simplification is the process of simplifying texts while preserving their meaning and information to increase understandability. Simplified sentences are easier to understand and easier to process by computers [4].

2.3.2 Applications and Approaches

Text simplification can be syntactic or lexical. To simplify texts lexically difficult words are substituted by easier words. For language learners and people with reading disability difficult words are less frequent words [11, 12, 20]. Hence one of the solutions is to replace those words by the most frequent synonym in their set of synonyms (including the word itself) [11, 20]. Since words can have different meanings, this approach often leads to meaningless sentences. To tackle this problem in lexical simplification, word sense disambiguation can be done [18]. For syntactic simplification, a set of rules [14, 18, 52] or log-linear models [4] can be utilized to simplify sentences by breaking down them into shorter and simpler sentences.

Text simplification has been studied for two main purposes: making text easier to understand for readers with aphasic disability [11] or low literacy skills [8] and as a preprocessing step for several NLP tasks such as relation extraction [33, 34], semantic role labeling [61], machine translation [46], summarization [53, 60], and improving accuracy of parsers [13, 14, 52]. Preprocessing text to simplify it has been inspired by the fact that performance of these systems rapidly deteriorates as the length and complexity of the sentence increases [13, 33]. Most of the errors in parsing are due to long, complex, and ambiguous sentences and it has been shown that text simplification eases summarization by shortening sentences and dropping non essential information. Silveira and Branco [53] showed that removing some specific structures such as relative clauses, explanatory phrases, and appositions does not decrease the readability and informativeness of the sentence and hence can be removed to simplify the sentences and output a better summery. Vanderwende et al. [60] proposed a better extractive summarization system by adding simplified sentences to the input so as to give the summarization system the option of choosing between simplified sentence and original sentence.

Unlike previous rule-based approaches for syntactic simplification which is limited to English language, in [4] a general framework has been proposed, TriS, for syntactic simplification by casting the problem into a search problem in which among all possible simplified sentences of each sentence, they find a subset of it that gives the highest probability given the original sentence e. In the other word, they find the subset S which maximizes the following equation:

$$\mathrm{p}(S|e) = \frac{\exp(\sum_{i=1}^{M} w_i \mathrm{f_i}(S, e))}{\sum_{S'} \exp(\sum_{i=1}^{M} w_i \mathrm{f_i}(S', e))} \tag{2.1}$$

In this equation, feature functions, *f(S,e)*, are based on 177 sentence level and interactive features extracted from original and simplified sentences. To learn weight of the features, w, they use online learner MIRA [17]. To build all the possible simplified sentences, they assume any simplified sentence has the following structure:

Subject +Verb+ Object

In which *Subject* and *Object* are noun phrase (NP). They make a list of all NPs in the original sentences (plus an empty NP for intransitive verbs) and a list of verbs in the original sentence. Then they make a list of all possible simple sentences by enumerating all the possible ways of combining these two lists. If there are n NPs and m verbs in the sentence this approach will yield n^2m simple sentences. At the end, they make use of stack decoding algorithms to find the best subset of this list of simplified sentences that maximize the equation 1 [4].

Jonnalagadda and Gonzalez [33] showed that syntactically simplifying sentences using a set of rules helps relation extraction in the domain of biomedical texts which usually have longer sentences with more abbreviations and relative clauses than less specialized and less technical texts like news. Being optimized to extract relations among proteins from biomedical scientific literature, it may not be very useful on other domains like conversational data. As opposed to scientific literature in which sentences are grammatically correct, sentences in conversational texts are not well-written. They are noise-prone and contain ungrammatical text with much cryptic content. But in both domains, more abbreviations than general text are usually used. As a consequence, we hypostatize that text simplification may be of benefit in the domain of conversational data as well. In chapter 4, we test this hypothesis through experiments and present results. In this study, TriS, has been used to simplify texts before feeding them into OLIIE.

Chapter 3

Methodology

To fairly evaluate Open IE over conversational data a test dataset covering different types of conversations and sentences is required. To the best of our knowledge there is no such dataset and hence we propose a method to create a dataset that has been sampled from a wide range of conversational corpora [9] including synchronous conversations (AMI and ICSI corpus), microblogs (tweets), threaded or asynchronous conversations (Email and blog threads), and reviews on products and services (Opinosis Dataset). In the next two sections, first the conversational corpora used for sampling has been described and then the method proposed for sampling sentences from these corpora has been described.

3.1 Dataset Creation

The test dataset used in this study includes a total of 600 sentences which were sampled from 6 conversational corpora (100 sentences from each). The sampling approach has been described in the next section.

The corpora cover a wide range of conversational data [9] including synchronous conversations (AMI and ICSI corpus), microblogs (tweets), threaded or asynchronous conversations (Email and blog threads), and reviews on products and services (Opinosis Dataset). Totally 6 corpora were used which have been described in the following sections:

3.1.1 Reviews

Writing review is a common way of expressing ideas and opinions about new products and services. They are usually informal and have colloquial language. Opinosis Dataset 1.0 was originally developed for summarization and contains reviews on 51 topics such as "battery life ipod nano 8gb" and "navigation amazon kindle". These reviews are about hotels, cars, and electronics and were collected from Tripadvisor, Edmunds.com and Amazon.com. [25]. This dataset was used as a representative of this conversational modality in our test dataset.

3.1.2 Emails

Writing and reading Emails has been the most popular conversational activity and hence one Email corpus was included in the evaluation of Open IE tools. BC3 Email dataset was originally developed for summarization as well and contains 40 email threads and 261 Emails from W3C corpus [59].

3.1.3 Meetings

Another important conversational modality is meeting. Many people spend a lot of time in meetings and due to advancement in transcribing, these spoken conversations now are available as conversational texts. We have used two different meeting corpora: AMI corpus consists of 100 hours of scenario and non-scenario meetings. We used the scenario portion of this corpus in which four persons participate in the meetings and talk about designing a remote control [10]; ICSI corpus consists of 75 non-scenario or natural technical meetings held by ICSI researchers [30]. Both corpora have native and none native English speakers but ICSI meetings has on average six to ten participants per meeting which is more than that of AMI scenario.

3.1.4 Blogs and Online Discussions

Blogs and forum discussions are another type of popular conversational texts in which people share their comments, thoughts and feelings about any topic posted by the first participant of the discussion which can be news, questions, events and so on. Slashdot is a website for news stories about technology along with lengthy discussions and comments of users. The dataset we used consists of all the threaded discussions of the users for 10 dates.

3.1.5 Social Networks

Nowadays people spend a great amount of time on updating their profile, reading their friends' posts, and commenting on them in social networks such as Twitter and Facebook. The language in social networks is informal and ungrammatical with lots of abbreviations. The dataset used as a representative of this type of conversations is 5146 random tweets taken from Twitter.

3.1.6 Dataset Characteristics

Characteristics of datasets has been shown in Table 3.1. As this table shows, On average Slashdot has longest senences while AMI has the shortest sen-

Name	#doc	#sent	#sent per doc	#word	#word per doc	#word per sent
ICSI	494	80410	162	839874	1700	10
Slashdot	15	8128	541	211180	14078	26
AMI	137	76865	556	716382	5191	9
Opinosis	51	6851	134	128150	2512	18
Twitter	5146	3254	813	90802	22700	10
BC3	40	2395	59	29642	741	12

Table 3.1: Dataset characteristics.

tences. For tokenizing words, the word tokenizer of NLTK toolkit was used. A stop list of special tokens were used to filter emoticons and tokens such as "LOL", "lolll", "ah", and "!!!". For Twitter dataset, urls and ReTweet (RT) expressions were removed from the tweets which increased the accuracy of extractions of OLIIE about 7% and that of OLIIE-Simplified about 20%. Because of conversational nature of these datasets word tokenizer made more errors and its accuracy were lower than other domains. In Table 3.1, the first column shows number of documents per dataset and the rest of columns show total number of sentences, average number of sentences per document, total number of words, average number of words per document, and average number of words per sentence in order. For twitter dataset each tweet, for BC3 corpus each thread, and for Opinosis each topic has been considered as one document.

3.1.7 Sampling Method

In order to evaluate the performance of a relation extraction tool, the test dataset ideally should contain different types of sentences having different types of relations. As a consequence, to obtain a representative sample of sentences, we used two stage stratified sampling. In the first stage, to capture key characteristics of each corpus, each corpus plays the role of one stratum independently. In the second stage, 100 sentences were sampled from each corpus (stratum). In the second stage, we did not use a simple stratified sampling. Instead, we extracted a set of syntactic and conversational features from each sentence and then we grouped them based on the resulting feature vectors. The stratified sampling has been done based on the probability of resulting groups. More members in the group, higher the probability to be chosen for sampling. The feature set used has been shown in Table 3.2.

The syntactic features are inspired by the fact that more punctuations and relative nouns in a sentence make it more challenging for relation extraction tools to extract relation from. Conversational features are those features which were found to be useful in the domain of conversational data. We chose a subset of features proposed by Murray and Carenini [44] that appeared to be useful in the relation extraction task. *SMT* shows the sum of *Tprob* scores. *Tprob* itself ishows the probability of each turn given the word. *CLOC* represents the sentence position in the conversation. Since it is often more difficult to extract relation from longer sentences, we utilized two features that employ the length of sentence: *SLEN* and *SLEN2* represent the number of words normalized by the longest sentence in the conversation and turn respectively. *CWS* shows conversation cohesion and is computed after removing stopwords. It represents the number of words appearing in other turns except the current turn. *CENT1* shows the similarity of the sentence to the conversation and is computed based on the cosine value between the sentence and the rest of the conversation.

Syntactic features	*Question*	A binary feature indicating whether the sentence is a question
	WH_count	Number of relative pronouns in the sentence
	Punc_count	Number of punctuations in the sentence
Conversational features	*SMT*	Sum of Tprob scores
	CLOC	Position in conversation
	SLEN	Globally normalized word count
	SLEN2	Locally normalized word count
	CWS	Rough ClueWordScore
	CENT1	Cosine of sentence and conv., w/ Sprob

Table 3.2: Feature set used in sampling sentences.

3.2 Open IE on Conversational Data

We were not able to use TreeKernel because of lack of training examples in our domain. SONEX and EXEMPLAR were not used since they extract relations between named entities and were able to extract only few relations while OLIIE extract hundreds of relations between noun phrases.

The relations extracted by OLIIE from the created dataset manually evaluated following a set of rules and they were labeled as correct if they were

adherent to those rules and deemed to be correct. OLIIE performance were evaluated before and after simplification based on the number of extracted relations, the accuracy of extracted arguments and relation phrases, and the informativeness of confidence score.

3.3 Text Simplification for Open IE

For the task of information extraction, only syntactic simplification will be useful since the purpose of lexical simplification is to improve readability for human not computer. Hence, in this study only the effect of syntactic simplification will be evaluated for the task of relation extraction. For text simplification, TriS which is a syntactic simplifier has been used to syntactically simplify texts before feeding them into OLLIE. Experimental results are presented and discussed in the next chapter.

Chapter 4

Experimental Results

4.1 Evaluation Metrics

Evaluating relation extraction approaches is difficult due to the subjectivity and ambiguity of the task. It is not only difficult for automatic systems but also for human to decide whether there is a relation in the sentence and if so, which words of the sentences form the relation [19, 63]. Another type of ambiguity arises from the definition of an entity. For example in the sentence "John went to Starbucks coffee shop", one may say the second argument of the relation "went" can be "Starbucks coffee shop", "Starbucks", or both, while someone else considers only "Starbucks coffee shop" as the correct argument.

Evaluating and comparing recall of Open IE tools becomes even more challenging than accuracy with the presence of implicit relations and relations for which the relation phrase has not been stated in the sentence. For example, in the sentence "Rumi, a poet, was born in Nishapur, Iran" one may consider the relation "located in" between Nishapur and Iran even though the relation phrase has not been appeared in the sentence. In the sentence "John broke the residence rules" one may conclude the implicit relation lives in between "John" and "the residence" as well. As a consequence, Open IE tools which extract relations between noun phrases don't report the recall of their system. Instead, they use accuracy and number of exactions as a way of comparison. As a consequence, the performance of OLLIE were evaluated before and after simplification based on the following metrics:

- The number of extracted relations
- The accuracy of extracted arguments and relation phrases
- Informativeness of confidence score

4.2 Results

Two experiments have been performed. In the first experiment, we fed the created dataset into OLLIE and report the accuracy of arguments and relation phrases extracted. In the second experiment, these sentences first were simplified by TriS and then they have been fed into OLLIE. The second system, OLLIE using text simplification as a preprocessing step, will be referred as OLIIE-Simplified. The result of these experiments has been shown in Table 4.1 and Table 4.2. In all tables, the bold numbres show the times OLLIE outperformed OLIIE-Simplified. The columns, from left to right, show number of extractions, accuracy of the first argument, accuracy of the first argument when relation phrase is correct, accuracy of the relation phrase, accuracy of the second argument, accuracy of the second argument when relation phrase is correct, and accuracy when both arguments and relation phrase are correct.

TriS failed to simplify most of the sentences in AMI and ICICS corpus due to lack of punctuations or wrong punctuations which has made sentences too long to be simplified by TriS. As a consequence, accuracy and average confidence score for them has not been reported in the second experiment. Whenever TriS did not have any suggestion for simplifying the sentence the original sentence were fed into OLIIE. As the Table 4.2 shows text simplification considerably improves accuracy of extractions for both arguments and relation in all cases except Slashdot dataset. TriS were not able to simplify sentences in Slashdot dataset correctly mostly because of errors in sentence tokenization.

Tables 4.3 and 4.4 show average confidence scores OLIIE and OLIIE-Simplified assigned to the extractions. From left to right, the columns show average confidence score of all extractions, average confidence score of correct relations, average confidence score of the incorrect relations, average confidence score when both the first argument and relation phrase is correct, average confidence score when the first argument is incorrect and relation is correct, average confidence score when both the second argument and relation phrase is correct, average confidence score when the second argument is incorrect and relation is correct, and average confidence score when both arguments and relation phrase are correct.

Figures 4.1 and 4.2 compares their accuracy and confidence scores when both arguments and relation phrase are correct. Figure 4.3 compares their confidence scores when relation phrase in incorrect. As this figure shows, OLLIE-Simplified assigned lower confidence scores to incorrect extractions on average.

Dataset	#Extractions	Arg1 acc.	Arg1 acc. when relation is correct	Relation phrase acc.	Arg2	Arg2 acc. when relation is correct	All correct acc.
ICSI	292	73.6%	56.8%	47.9%	66.8%	57.9%	45.2%
AMI	650	80.0%	61.2%	71.5%	66.3%	52.0%	43.2%
BC3	**148**	**79.0%**	61.5%	73.0%	69.6%	32.4%	48.6%
Slashdot	**301**	**79.5%**	**65.4%**	76.4%	74.3%	**63.7%**	**54.1%**
Reviews	**372**	**65.6%**	51.3%	64.5%	61.8%	53.5%	40.9%
Twitter	90	66.7%	55.6%	62.2%	70.0%	52.2%	45.6%

Table 4.1: Accuracy before simplification. The bold numbres show the cases OLLIE outperformed OLIIE-Simplified.

Dataset	#Extractions	Arg1 acc.	Arg1 acc. when relation is correct	Relation phrase acc.	Arg2 acc.	Arg2 acc. when relation is correct	All correct acc.
BC3	141	74.5%	**66.0%**	**77.3%**	**71.6%**	**68.1%**	**58.2%**
Slashdot	211	77.3%	63.6%	**76.8%**	**74.7%**	63.6%	51.5%
Reviews	233	65.2%	**55.4%**	**68.2%**	**65.7%**	**54.9%**	**44.2%**
Twitter	**99**	**73.7%**	**63.6%**	**72.7%**	**79.8%**	**64.6%**	**55.6%**

Table 4.2: Accuracy after simplification. The bold numbres show the cases OLLIE-Simplified outperformed OLIIE.

Dataset	All ext	Corr. rel	Incorr. rel	Corr. Arg1 and rel	Incorr. Arg1 and corr. rel	Corr. Arg2 and rel	Incorr. Arg2 and correct rel	All corr.
ICSI	0.6	0.43	0.15	0.36	0.07	0.36	0.07	0.29
AMI	0.56	0.4	0.15	0.35	0.05	0.3	0.09	0.26
BC3	0.61	0.46	0.14	0.39	**0.06**	0.37	0.09	0.31
Slashdot	0.66	0.49	0.14	**0.42**	**0.07**	0.41	0.08	**0.35**
Reviews	0.64	0.42	0.21	0.34	0.08	0.34	**0.08**	0.27
Twitter	0.7	0.44	0.26	0.4	**0.04**	0.38	0.06	0.34

Table 4.3: Average confidence score before simplification. The bold numbres show the cases OLLIE outperformed OLIIE-Simplified.

Dataset	All ext	Corr. rel	Incorr. rel	Corr. Arg1 and rel	Incorr. Arg1 and corr. rel	Corr. Arg2 and rel	Incorr. Arg2 and correct rel	All corr.
BC3	**0.69**	**0.54**	**0.11**	**0.46**	0.08	**0.48**	**0.06**	**0.4**
Slashdot	**0.7**	0.49	**0.12**	0.4	0.09	0.41	0.08	0.33
Reviews	**0.66**	**0.47**	**0.17**	**0.39**	0.08	**0.38**	0.09	**0.31**
Twitter	**0.74**	**0.54**	**0.19**	**0.47**	0.07	**0.48**	**0.05**	**0.42**

Table 4.4: Average confidence score after simplification. The bold numbres show the cases OLLIE-Simplified outperformed OLIIE.

4.3 Analysis and Discussion

If we analyze the accuracies of relation phrase and arguments when relation phrase is correct, we see that OLLIE has the best performance on Slashdot and BC3 corpora in order and the worst on Reviews corpus. As Table 4.2 shows OLLIE-Simplified has also the worst performance on Reviews corpus but the best performance on BC3 corpus. Both systems have the worst performance on Reviews corpus which might be due to language of reviews. In reviews, people express their opinions and feelings using phrases and incomplete sentences. Some examples of such sentences are as follows: "accurate GPS for not so much money", "Better battery life", "FAR better with wireless function on", "NO USER REPLACEABLE BATTERY", "Easy to read, navigate, etc.", "Significant improvements to ergonomics and navigation". With the same logic, better performance on Slashdot and BC3 corpora might be due to language in these corpora. BC3 corpus contains emails written in a corporation which usually have more formal and grammatical sentences. Slashdot is a website for news stories about technology along with lengthy discussions and comments of users with probably more technical and grammatical content. Overall, the most difficult conversational modality for relation extraction for both systems is reviews. The easiest ones for OLIIE are blogs and emails and for OLLIE-Simplified are emails and microblogs (Twitter).

According to Table 4.1 and 4.2, both systems extract more relations from Review corpus[8]. OLLIE extracts more relations than OLLIE-Simplified except for Twitter corpus for which OLLIE-Simplified extracts more. It might be due to the way OLLIE works. OLLIE tries the same relation phrase with different pairs of arguments but only one of these extractions is correct. Even though OLLIE-Simplified extracts fewer number of extractions for most corpora, the extractions are more distinct.

As Table 4.3 and 4.4 show, OLLIE is less confident in AMI and ICSI extractions while OLLIE-Simplified is less confident in Reviews extractions. Considering all extractions regardless of being correct or not, both OLLIE and OLLIE-Simplified are most confident in Twitter extractions.

As results show, OLLIE-Simplified assigned lower confidence scores to incorrect relation phrases in all cases and higher confidence scores to correct extractions for most datasets. As a consequence we conclude that text simplification improves the informativeness of confidence scores.

[8] As we were not able to run TriS on AMI and ICSI corpora we omit them here in comparison.

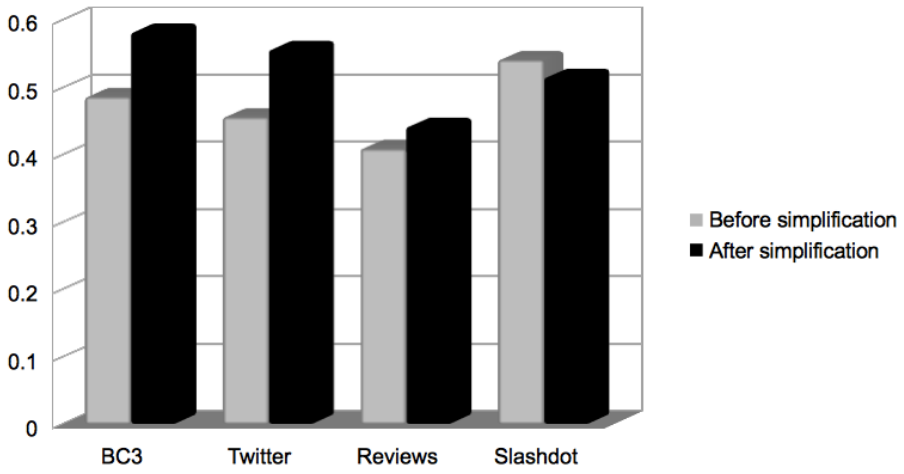

Figure 4.1: Accuracy of extraction when the both arguments and relation phrase are correct. The largest increase in the accuracy can be seen for BC3 and Twitter corpus, 13% and 10% in order.

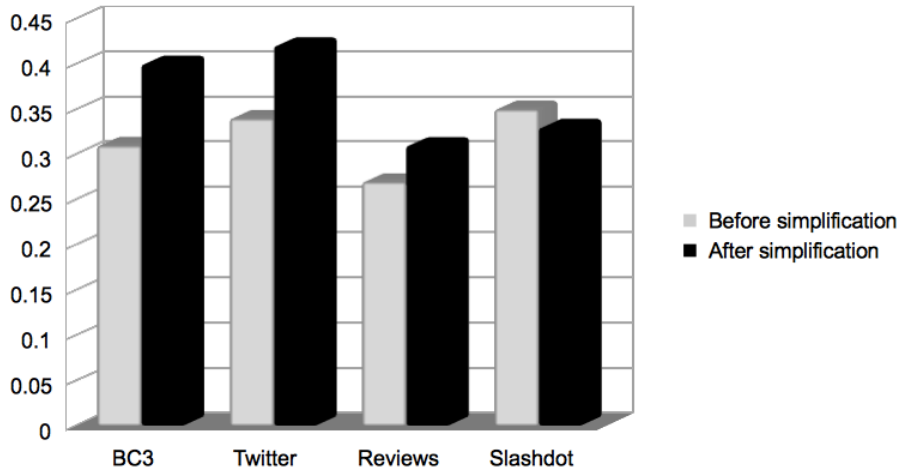

Figure 4.2: Average confidence score when the both arguments and relation phrase are correct. The largest increase in the confidence score can be seen for BC3 and Twitter corpus in order.

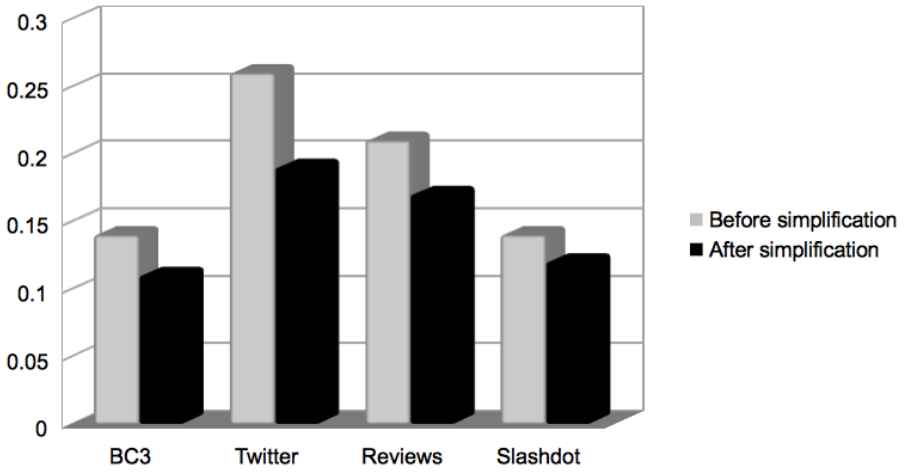

Figure 4.3: Average confidence score when the relation phrase is incorrect. The largest decrease in the confidence score of incorrect relation phrases has happend for Twitter corpus.

As figures 4.1 and 4.2 show the largest increase in accuracy and confidence score can be seen for BC3 and Twitter corpora in order. As opposed to what we thought text simplification has been much more useful and advantageous for corpora with shorter sentences. As figure 4.3 shows the largest decrease in confidence score of incorrect relation phrases has happend for Twitter corpus.

Text simplification is very effective in increasing accuracy of OLLIE for Twitter dataset. Tweets have much more cryptic content and abbreviations than other conversations since they must be at most 140 characters long. This results again verifies the result found by Jonnalagadda and Gonzalez [33] that text simplification greatly helps relation extraction when the amount of cryptic content and abbreviations in text is much more than less specialized and less technical texts like news.

Text simplification is not very effective in increasing the accuracy of OLLIE for Slashdot dataset due to errors in sentence tokenization and its lengthy sentences. Sentence tokenizer made more mistakes in this dataset than other datasets which lead to poor performance of TriS in simplifying sentences in Slashdot. Slashdot has the longest sentences among other corpora and as opposed to what we thought, TriS did not work well when

sentences were too long. It might be due to the way TriS simplifies sentences in which it builds its search space based on all noun phrases in the sentences.

Each system finds distinct relations not found by the other system. Hence a new system which utilizes the union of extracted relations of the two systems will outperform both systems in terms of recall.

Another interesting finding is that OLLIE is more capable in accurately finding the first argument of the relation while OLLIE-Simplified more accurately extracts the relation phrase and the second argument.

Chapter 5

Conclusion and Future Work

To evaluate Open IE in the domain of conversational texts, a method was proposed to create a test dataset covering a wide range of conversational data.

Conversational text poses new challenges due to its specific characteristics including cryptic content, lots of abbreviations, ungrammatical and informal language. As a consequence text simplification was used to mitigate the problems.

We discussed why text simplification will be useful for this task and should be used as a preprocessing step in relation extraction. The approach taken to sample from conversational datasets and experiments were described and two systems were compared.

To the best of our knowledge, this is the first time Open IE has been evaluated in the domain of conversational data. We proposed a method to sample a test dataset covering a wide range of conversational data. We showed text simplification empowers relation extraction in the domain of conversational texts. Experimental results show that OLLIE-Simplified outperforms OLLIE in terms of accuracy and informativeness of the confidence score.

As opposed to what we hypothesized text simplification has been much more useful and advantageous for corpora with shorter average sentences. Text simplification has been much more effective in increasing the accuracy of OLLIE for Twitter dataset which has much more cryptic content and abbreviations than other conversations due to its length limit (at most 140 characters).

Overall, the most difficult conversational modality for relation extraction for both systems is reviews. The easiest ones for OLIIE are blogs and emails and for OLLIE-Simplified are emails and microblogs (Twitter). In reviews, people express their opinions and feelings using phrases and incomplete sentences leading to difficulty in relation extraction. Emails written in a corporation usually have more formal and grammatical sentences and technical blogs like Slashdot have more technical and grammatical sentences which helps relation extraction.

Each system finds distinct relations not found by the other system. OLLIE-Simplified can find new relations not already found by OLLIE and hence a new system which utilizes the union of extracted relations of two systems will outperform both systems in terms of recall. OLLIE is more capable in accurately finding the first argument of the relation while OLLIE-Simplified more accurately extracts the relation phrase and the second argument. A unified system that takes advantage of these findings, would outperform both systems.

Since conversational data has special characteristics, a text simplifier developed to deal with conversational data would be of more benefit. Another direction will be to evaluate the effect of text simplification on other Open IE systems and in other domains.

Bibliography

[1] Eugene Agichtein. Scaling information extraction to large document collections. *IEEE Data Eng. Bull*, 28:3–10, 2005.

[2] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA, 2000. ACM.

[3] Kemafor Anyanwu, Angela Maduko, and Amit Sheth. Semrank: Ranking complex relationship search results on the semantic web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 117–127, New York, NY, USA, 2005. ACM.

[4] Nguyen Bach, Qin Gao, Stephan Vogel, and Alex Waibel. Tris: A statistical sentence simplifier with log-linear models and margin-based discriminative training. In *IJCNLP*, pages 474–482, 2011.

[5] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2670–2676, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[6] Michele Banko and Oren Etzioni. The tradeoffs between open and traditional relation extraction.

[7] Sergey Brin, Rajeev Motwani, Lawrence Page, and Terry Winograd. What can you do with a web in your pocket? *IEEE Data Eng. Bull.*, 21(2):37–47, 1998.

[8] Arnaldo Candido, Jr., Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio. Supporting the adaptation of texts for poor literacy readers: A text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on*

Innovative Use of NLP for Building Educational Applications, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[9] Giuseppe Carenini, Gabriel Murray, and Raymond Ng. *Methods for Mining and Summarizing Text Conversations*. Morgan & Claypool Publishers, 1st edition, 2011.

[10] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 28–39, Berlin, Heidelberg, 2006. Springer-Verlag.

[11] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. In *In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, 1998.

[12] John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. Simplifying text for language-impaired readers. In *In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL*, pages 269–270, 1999.

[13] R. Chandrasekar, Christine Doran, and B. Srinivas. Motivations and methods for text simplification. In *PROCEEDINGS OF THE SIXTEENTH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING '96)*, 1996.

[14] R. Chandrasekar and B. Srinivas. Automatic induction of rules for text simplification, 1997.

[15] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the Sixth International Conference on Knowledge Capture*, K-CAP '11, pages 113–120, New York, NY, USA, 2011. ACM.

[16] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun ichi Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In Russ B. Altman, Tiffany Murray, Teri E. Klein, A. Keith Dunker, and Lawrence Hunter, editors, *Pacific Symposium on Biocomputing*, pages 4–15. World Scientific, 2006.

[17] Koby Crammer and Yoram Singer. Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res.*, 3:951–991, March 2003.

[18] Jan De Belder and Marie-Francine Moens. Text simplification for children. In *Prroceedings of the SIGIR Workshop on Accessible Search Systems,*, pages 19–26. ACM, 2010.

[19] Filipe de S Mesquita, Jordan Schmidek, and Denilson Barbosa. Effectiveness and efficiency of open relation extraction. In *EMNLP*, pages 447–457. ACL, 2013.

[20] Siobhan Devlin and Gary Unthank. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '06, pages 225–226, New York, NY, USA, 2006. ACM.

[21] Doug Downey, Stefan Schoenmackers, and Oren Etzioni. Sparse information extraction: Unsupervised language models to the rescue. In *In Proc. of ACL*, pages 696–703, 2007.

[22] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *ARTIFICIAL INTELLIGENCE*, 165:91–134, 2005.

[23] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 3–10. AAAI Press, 2011.

[24] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[25] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 340–348, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[26] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[27] Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 286–295, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[28] Hui Huang, Xiaogang Wu, Ragini Pandey, Jiao Li, Guoling Zhao, Sara Ibrahim, and Jake Y Chen. C2maps: a network pharmacology database with comprehensive disease-gene-drug connectivity relationships. *BMC Genomics*, 13 Suppl 6:S17, 2012.

[29] Minlie Huang and et al. Discovering patterns to extract protein-protein interactions from full texts, 2004.

[30] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The icsi meeting corpus. pages 364–367, 2003.

[31] Nitin Jindal and Bing Liu. Review spam detection. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1189–1190, New York, NY, USA, 2007. ACM.

[32] Richard Johansson and Pierre Nugues. Dependency-based semantic role labeling of propbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 69–78, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[33] Siddhartha Jonnalagadda and Graciela Gonzalez. Biosimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. *CoRR*, abs/1107.5744, 2011.

[34] Siddhartha Jonnalagadda, Luis Tari, Jörg Hakenberg, Chitta Baral, and Graciela Gonzalez. Towards effective sentence simplification for automatic processing of biomedical text. *CoRR*, abs/1001.4277, 2010.

[35] Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. Topic segmentation and labeling in asynchronous conversations. *J. Artif. Int. Res.*, 47(1):521–573, May 2013.

[36] Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*. ACM, 2004.

[37] Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *CEAS*, 2004.

[38] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 523–534, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[39] Diana Maynard, Adam Funk, and Wim Peters. Sprat: a tool for automatic semantic pattern-based ontology population. In *IN: INTERNATIONAL CONFERENCE FOR DIGITAL LIBRARIES AND THE SEMANTIC WEB*, 2009.

[40] Ryan Mcdonald, Fernando Pereira, Seth Kulick, Scott Winters, Yang Jin, and Pete White. Simple algorithms for complex relation extraction with applications to biomedical ie. In *In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05*, pages 491–498, 2005.

[41] Yashar Mehdad, Giuseppe Carenini, Frank Tompa, and Raymond T. NG. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[42] Yuval Merhav, Filipe Mesquita, Denilson Barbosa, Wai Gen Yee, and Ophir Frieder. Extracting information networks from the blogosphere. *ACM Trans. Web*, 6(3):11:1–11:33, October 2012.

[43] Junichiro Mori, Takumi Tsujishita, Yutaka Matsuo, and Mitsuru Ishizuka. Extracting relations in social networks from the web using similarity between collective contexts. In *International Semantic Web Conference*, pages 487–500, 2006.

[44] Gabriel Murray and Giuseppe Carenini. Summarizing spoken and written conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 773–782, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[45] Gabriel Murray and Giuseppe Carenini. Subjectivity detection in spoken and written conversations. *Nat. Lang. Eng.*, 17(3):397–418, July 2011.

[46] Francisco Oliveira, Fai Wong, and Iok-Sai Hong. Systematic processing of long sentences in rule based portuguese-chinese machine translation. In *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'10, pages 417–426, Berlin, Heidelberg, 2010. Springer-Verlag.

[47] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[48] Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 1–10, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[49] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[50] Benjamin Rozenfeld and Ronen Feldman. Self-supervised relation extraction from the web. *Knowl. Inf. Syst.*, 17(1):17–33, October 2008.

[51] Yue Shang, Yanpeng Li, Hongfei Lin, and Zhihao Yang. Enhancing biomedical text summarization using semantic relation extraction, 2011.

[52] Advaith Siddharthan. Syntactic simplification and text cohesion. Technical report, Research on Language and Computation, 2003.

[53] Sara Silveira and Antnio Branco. Enhancing multi-document summaries with sentence simplification. In *Proceedings of International Conference on Artificial Intelligence*, Las Vegas, USA, July 2012.

[54] Fabian M. Suchanek. Leila: Learning to extract information by linguistic analysis. In *In Workshop on Ontology Population at ACL/COLING*, pages 18–25, 2006.

[55] Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. Sofie: A self-organizing framework for information extraction. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 631–640, New York, NY, USA, 2009. ACM.

[56] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 8–15, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[57] Maryam Tavafi, Yashar Mehdad, Shafiq Joty, Giuseppe Carenini, and Raymond Ng. Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 117–121, Metz, France, August 2013. Association for Computational Linguistics.

[58] Ivan Titov and Alexandre Klementiev. A bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1445–1455, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[59] J. Ulrich, G. Murray, and G. Carenini. A publicly available annotated corpus for supervised email summarization. In *AAAI08 EMAIL Workshop*, Chicago, USA, 2008. AAAI.

[60] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*, 43(6):1606–1618, November 2007.

[61] David Vickrey and Daphne Koller. Sentence simplification for semantic role labeling, 2008.

[62] Fei Wu and Daniel S. Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[63] Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. Open information extraction with tree kernels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 868–877, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[64] Alexander Yates and Oren Etzioni. Unsupervised resolution of objects and relations on the web. In *Proceedings of NAACL HLT*, pages 121–130, Rochester, NY, April 2007.

[65] Haijun Zhai, Todd Lingren, Louise Deleger, Qi Li, Megan Kaiser, Laura Stoutenborough, and Imre Solti. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processin. *J Med Internet Res*, 15(4), April 2013.

[66] Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou. A composite kernel to extract relations between entities with both flat and structured features. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 825–832, Morristown, NJ, USA, 2006. Association for Computational Linguistics.