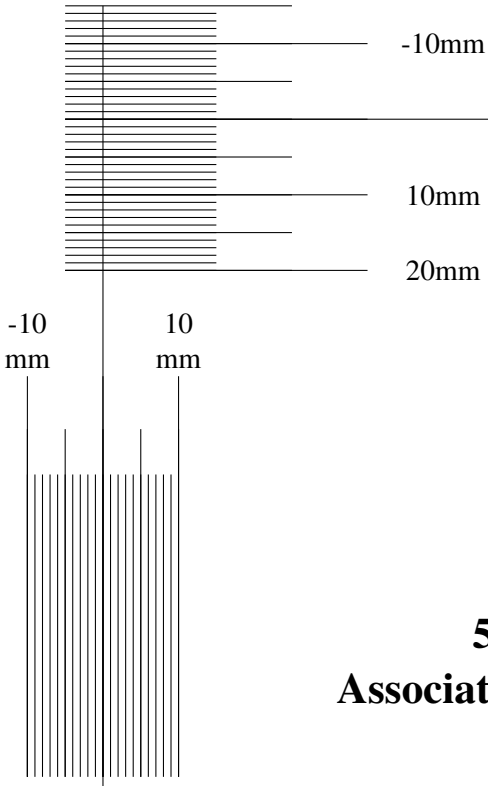


ACL 2012

50th Annual Meeting of the Association for Computational Linguistics

Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP

July 8 - 14, 2012
Jeju, Republic of Korea

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-937284-36-7

Introduction

Recent emergence of Collaboratively Constructed Semantic Resources has led to a set of new research topics that became visible at major computational linguistics conferences. Researchers have studied methods to utilize collaboratively constructed resources to substitute or supplement conventional lexical semantic and linguistically annotated resources. At the same time, Natural Language Processing techniques have been applied to enhance the collaboration process and the overall quality of the collaboratively constructed resources.

This volume contains papers accepted for presentation at the 3rd Workshop on Collaboratively Constructed Semantic Resources and their Applications to NLP, which took place on July 13, 2012, as part of the ACL 2012 conference in Jeju, Republic of Korea. We received submissions from a broad spectrum of research topics, going beyond the coverage of the previous workshops. After careful review by our program committee, three long papers and three short papers were finally accepted for presentation.

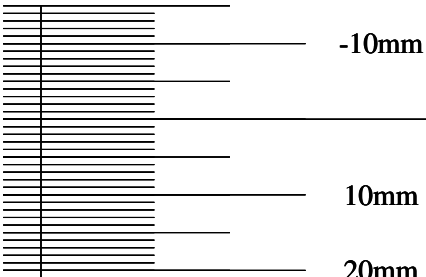
The preceding "People's Web meets NLP" workshops at ACL-IJCNLP 2009 and COLING 2010 have successfully gathered researchers from different areas and allowed an interdisciplinary exchange of research outcomes and ideas. The workshop has contributed to the creation of valuable semantic resources and tools based on Collaboratively Constructed Semantic Resources and the consolidation of this young research field.

As the third one in the series, this workshop aims to intensify research that demonstrates the effectiveness of the resources mined from Collaboratively Constructed Semantic Resources in diverse Natural Language Processing tasks. We issued a call for both long and short papers. We especially encouraged submissions that show the benefit of Collaboratively Constructed Semantic Resources in diverse Natural Language Processing tasks.

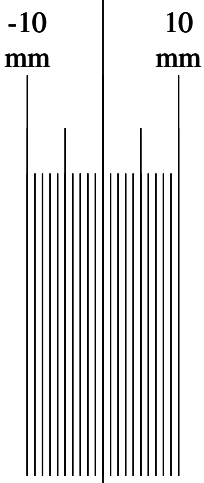
The aim of the People's Web Meets NLP workshop series is to bring together researchers with different backgrounds: from computational linguistics to various Natural Language Processing areas that benefit from collaboratively constructed semantic resources. We hope this is well reflected in the proceedings.

Jeju, July 2012

Iryna Gurevych, Nicoletta Calzolari Zamorani, and Jungi Kim



Paper ID 0



Organizers: 20mm

-10 mm 10 mm

Iryna Gurevych, Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt
Nicoletta Calzolari Zamorani, Istituto di Linguistica Computazionale, CNR
Jungi Kim, Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt

Program Committee:

Andras Csomai, Google Inc.
Andreas Hotho, Julius-Maximilians-Universität Würzburg
Anette Frank, Heidelberg University
Benno Stein, Bauhaus University Weimar
Christian M. Meyer, Technische Universität Darmstadt
David Milne, University of Waikato
Delphine Bernhard, University of Strasbourg
Diana McCarthy, Lexical Computing Ltd, UK
Donald Metzler, Information Sciences Institute, University of Southern California
Emily Pitler, University of Pennsylvania
Ernesto William De Luca, Technische Universität Berlin
Florian Laws, University of Stuttgart
Gerard de Melo, UC Berkeley
German Rigau, University of the Basque Country
Graeme Hirst, University of Toronto
Günter Neumann, DFKI Saarbrücken
Ido Dagan, Bar Ilan University
John McCrae, University of Bielefeld
Jong-Hyeok Lee, Pohang University of Science and Technology
Judith Eckle-Kohler, Technische Universität Darmstadt
Magnus Sahlgren, Swedish Institute of Computer Science
Manfred Stede, Universität Potsdam
Massimo Poesio, University of Essex
Omar Alonso, Microsoft Bing
Paul Buitelaar, DERI, National University of Ireland, Galway
Rene Witte, Concordia University Montréal
Roxana Girju, University of Illinois at Urbana-Champaign
Saif Mohammad, National Research Council Canada
Shuming Shi, Microsoft Research
Sören Auer, Leipzig University
Tat-Seng Chua, National University of Singapore
Tonio Wandmacher, SYSTRAN, Paris, France
Zornitsa Kozareva, Information Sciences Institute, University of Southern California

Invited Speaker:

Massimo Poesio, University of Essex

Title: Phrase Detectives: the First Three Years

Abstract: Phrase Detectives, one of the first games-with-a-purpose for corpus annotation (`www.phrasedetectives.org`) went officially online on December 1st 2008, and one of its very first presentations in front of an NLP audience took place at the first edition of the "People's Web Meets NLP" workshop in Singapore in 2009. The option of annotating Italian documents was added in 2010, and a Facebook version went live in January 2012. Although the project that funded its creation ended in September 2009, the game has stayed very much alive, in fact it is getting more active all the time - we recently passed the 11,000 players mark and are about to reach 200,000 words of fully annotated documents, with a goal of annotating at least 1 million. In the talk I will discuss recent developments and analyze the results so far in terms of quality and quantity of annotated data and annotation costs.

References: Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi, In Press - Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation, ACM Transactions on Interactive Intelligent Systems.

Short bio: Massimo Poesio is a Professor of Computer Science at the University of Essex and the Director of the Language Interaction and Computation Lab at the Centre for Mind/Brain Sciences, University of Trento. He is best known for his work in anaphora resolution, corpus annotation, and the acquisition of common sense knowledge.

Table of Contents

Sentiment Analysis Using a Novel Human Computation Game
Claudiu Cristian Musat, Alireza Ghasemi and Boi Faltings 1

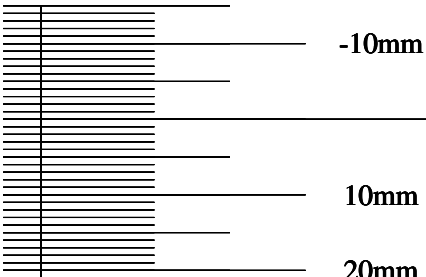
A Serious Game for Building a Portuguese Lexical-Semantic Network
Mathieu Mangeot and Carlos Ramisch 10

Collaboratively Building Language Resources while Localising the Web
Asanka Wasala, Reinhard Schäler, Ruvan Weerasinghe and Chris Exton 15

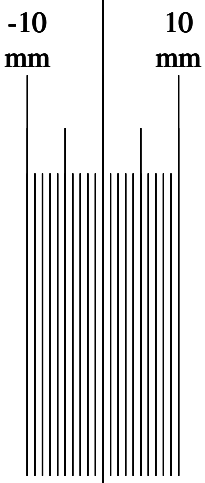
Resolving Task Specification and Path Inconsistency in Taxonomy Construction
Hui Yang 20

EAGER: Extending Automatically Gazetteers for Entity Recognition
Omer Farukhan Gunes, Tim Furche, Christian Schallhart, Jens Lehmann and Axel-Cyrille Ngonga Ngomo 29

Extracting Context-Rich Entailment Rules from Wikipedia Revision History
Elena Cabrio, Bernardo Magnini and Angelina Ivanova 34



Paper ID 0



Workshop Program

Friday July 13, 2012

09:10–09:30 Opening remarks

09:35–10:05 *Sentiment Analysis Using a Novel Human Computation Game*
Claudiu Cristian Musat, Alireza Ghasemi and Boi Faltings

10:10–10:30 *A Serious Game for Building a Portuguese Lexical-Semantic Network*
Mathieu Mangeot and Carlos Ramisch

10:30–11:00 Coffee break

11:00–11:20 *Collaboratively Building Language Resources while Localising the Web*
Asanka Wasala, Reinhard Schäler, Ruvan Weerasinghe and Chris Exton

11:25–12:30 Invited talk: Phrase Detectives: The First Three Years by Massimo Poesio

12:30–14:00 Lunch break

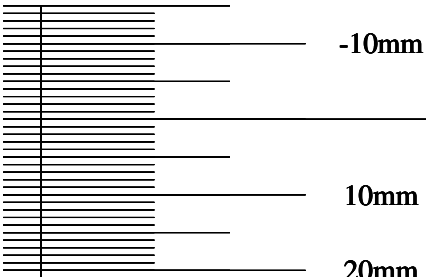
14:00–14:30 *Resolving Task Specification and Path Inconsistency in Taxonomy Construction*
Hui Yang

14:35–15:55 *EAGER: Extending Automatically Gazetteers for Entity Recognition*
Omer Farukhan Gunes, Tim Furche, Christian Schallhart, Jens Lehmann and Axel-Cyrille Ngonga Ngomo

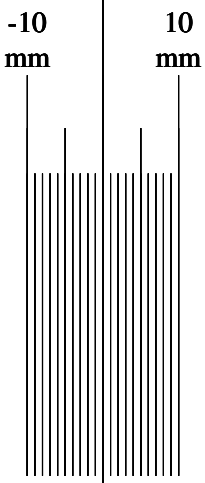
15:00–15:30 *Extracting Context-Rich Entailment Rules from Wikipedia Revision History*
Elena Cabrio, Bernardo Magnini and Angelina Ivanova

15:30–16:00 Coffee break

16:00–17:00 Panel discussion: Collaboratively Looking Ahead: How to Make Sustainable Goods out of Collaboratively Constructed Semantic Resources?



Paper ID 0



Sentiment Analysis Using a Novel Human Computation Game

Claudiu-Cristian Musat, THISONE **Alireza Ghasemi** **Boi Faltings**

Artificial Intelligence Laboratory (LIA)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
IN-Ecublens, 1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

Abstract

In this paper, we propose a novel human computation game for sentiment analysis. Our game aims at annotating sentiments of a collection of text documents and simultaneously constructing a highly discriminative lexicon of positive and negative phrases.

Human computation games have been widely used in recent years to acquire human knowledge and use it to solve problems which are infeasible to solve by machine intelligence. We package the problems of lexicon construction and sentiment detection as a single human computation game. We compare the results obtained by the game with that of other well-known sentiment detection approaches. Obtained results are promising and show improvements over traditional approaches.

1 Introduction

We propose a novel solution for the analysis of sentiment expressed in text media. Novel corpus based and lexicon based sentiment analysis methods are created each year. The continual emergence of conceptually similar methods for this known problem shows that a satisfactory solution has still not been found. We believe that the lack of suitable labelled data that could be used in machine learning techniques to train sentiment classifiers is one of the major reasons the field of sentiment analysis is not advancing more rapidly.

Recognizing that knowledge for understanding sentiment is common sense and does not require experts, we plan to take a new approach where labelled data is obtained from people using human computation platforms and games. We also prove that the method can provide not only labelled texts, but people also help by selecting sentiment-expressing features that can generalize well.

Human computation is a newly emerging paradigm. It tries to solve large-scale problems by utilizing human knowledge and has proven useful in solving various problems (Von Ahn and Dabbish, 2004; Von Ahn, 2006; Von Ahn et al., 2006a).

To obtain high quality solution from human computation, people should be motivated to make their best effort. One way to incentivize people for submitting high-quality results is to package the problem at hand as a game and request people to play it. This process is called gamification. The game design should be such that the solution to the main problems can be formed by appropriately aggregating results of played games.

In this work, we propose a cooperative human computation game for sentiment analysis called Guesstiment. It aims at annotating sentiment of a collection of text documents, and simultaneously constructing a lexicon of highly polarized (positive and negative) words which can further be used for sentiment detection tasks. By playing a collaborative game, people rate hotel reviews as positive and negative and select words and phrases within the reviews that best express the chosen polarity.

We compare these annotations with those obtained during a former crowd-sourcing survey and prove that packaging the problem as a game can improve the quality of the responses. We also compare our approach with the state-of-the-art machine

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 1–9,
Jeju, Republic of Korea, 8-14 July 2012.

learning techniques and prove the superiority of human cognition for this task. In a third experiment we use the same annotations in a multi faceted opinion classification problem and find that results are superior to those obtained using known linguistic resources.

In (section 2) we review the literature related to our work. We then outline the game and its rules (section 3). We compare the Guesstiment results to the state-of-the-art machine learning, standard crowd-sourcing methods and sentiment dictionaries(section 4) and conclude the paper with ideas for future work (section 5).

2 Related Work

In this section we review the important literature related and similar to our work. Sine we propose a human computation approach for sentiment analysis, we start by reviewing the literature on human computation and the closely related field of crowd-sourcing. Then we move on by having a brief look on the human computation and knowledge acquisition games proposed so far by the researchers. Finally, we briefly review major sentiment analysis methods utilized by the researchers.

2.1 Human Computation and Crowd-Sourcing

The literature on human computation is highly overlapping with that of crowd-sourcing, as they are closely connected. The two terms are sometimes used interchangeably although they are slightly different. Crowd-sourcing in its broadest form, "is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call"(Quinn and Bederson, 2011; Howe, 2006). Since the first use of the word crowd-sourcing by J. Howe (Howe, 2006), there has been a lot of interest in this field due to the wide accessibility of anonymous crowd workers across the web.

The work described in (Rumshisky, 2011) uses crowd-sourcing to perform word sense disambiguation on a corpus. In (Vondrick et al., 2010), crowd-sourcing is used for video annotation. Moreover, (Christophe et al., 2010) has used crowd-sourcing for satellite image analysis.

(Settles, 2011a) is another approach which aims at combining active learning with crowd-sourcing for text classification. The principal contribution of their work is that as well as document annotation, they use human computation also to perform feature selection. (Law et al.,) is another recent work which proposes a game for acquisition of attribute-value pairs from images.

2.2 Human Computation Games

Luis Von Ahn, the pioneer of the field of human computation, designed a game to encourage players to semantically annotate a large corpus of images (Von Ahn and Dabbish, 2004). It was the first human computation game.

Following Von Ahn's work, more researchers were encouraged to package computational problems as joyful games and have a group of non-expert users play them (Von Ahn, 2006). Verbosity (Von Ahn et al., 2006a) was designed with the goal of gathering common sense knowledge about words. KissKissBan (Ho et al., 2009) was another game for image annotation.

Peekaboom(Von Ahn et al., 2006b) aimed at image segmentation and "Phrase Detectives" (Chamberlain et al., 2008) was used to help constructing an anamorphic corpus for NLP tasks. Another human computation game is described in (Riek et al., 2011) whose purpose is semantic annotation of video data.

2.3 Sentiment Analysis

The field of sentiment analysis and classification currently mostly deals with extracting sentiment from text data. Various methods (Turney, 2002; Esuli and Sebastiani, 2006; Taboada et al., 2011; Pang et al., 2002) have been proposed for effective and efficient sentiment extraction of large collections of text documents.

Sentiment classification methods are usually divided into two main categories: Lexicon based techniques and methods based on machine learning. In lexicon-based methods, a rich lexicon of polarized words is used to find key sentences and phrases in text documents which can be used to describe sentiment of the whole text (Taboada et al., 2011). Machine learning methods, on the other hand, treat the sentiment detection problem as a text classification task (Pang et al., 2002).

A Serious Game for Building a Portuguese Lexical-Semantic Network

10mm
20mm
Mathieu Mangeot[♠] **Carlos Ramisch**[♠♣]
[♠] GETALP — LIG, University of Grenoble (France)
[♣] Federal University of Rio Grande do Sul (Brazil)
FirstName.LastName@imag.fr

-10 mm 10 mm

Abstract

This paper presents a game with a purpose for the construction of a Portuguese lexical-semantic network. The network creation is implicit, as players collaboratively create links between words while they have fun. We describe the principles and implementation of the platform. As this is an ongoing project, we discuss challenges and long-term goals.We present the current network in terms a quantitative and qualitative analysis, comparing it to other resources. Finally, we describe our target applications.

1 Introduction

The creation of lexical resources like wordnets is time consuming and very costly in terms of manpower. Funding agencies and publishing houses are very reluctant to launch new projects. Ironically, in our globalized nowadays world, the need of such resources for communication is growing. In this context, there is hope for building resources via communities of voluntary contributors. But is it possible to use the Wikipedia paradigm for building a rich and broad-coverage lexical resource reusable by humans and machines in NLP projects? Wordnets are very interesting resources, but they suffer of several limitations. First, even if the English wordnet (Miller et al., 1990) is open source and freely available, this is not the case of the EuroWordnets. Second, wordnets and other manually built thesauri are based on linguists' intuition. Information about up-to-date entities (Facebook, Costa Concordia, etc.) and real-world facts are missing. Third, relations between the synsets of wordnets are of limited semantic kinds. We would like to build other relations at the syntactic and lexical level (e.g. collocations).

Our first goal is to build a rich lexical network for the Portuguese language. The relations between nodes (words) is represented in a sophisticated way, by using lexical-semantical functions à la Mel'čuk (Mel'čuk, 1995) such as the Magn function representing the notion of intensifier: Magn(smoker) = heavy, Magn(bachelor) = confirmed. The resulting network represents the usage of the language, not the norm. Thus, it may contain frequent spelling mistakes or neologisms. This resource is open-source and freely available. It can be used in several applications: lexicography, printed dictionary, text generation, semantic information extraction, ontology learning, etc. The construction of the resource is done indirectly by contributors through a game.

In the next section, the concept of using serious games for building NLP resources will be explained (§ 2). The following section will detail the construction of the Portuguese version of the game (§ 3). Afterwards, we will discuss some preliminary results (§ 4) and finally we present future work (§ 5).

2 Serious Games and NLP

The concept of human contribution, collaboration and computation has been utilized in many applications and scenarios. The work of Luis von Ahn made a breakthrough, especially in ESP game (von Ahn, 2006; von Ahn and Dabbish, 2008). Human computation (crowdsourcing, volunteer contribution) is now seriously considered to be able to solve large computational problems (Speer, 2007). The idea of collecting massive contributions from volunteers through an online game took off recently. Nowa-

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 10–14,
Jeju, Republic of Korea, 8-14 July 2012.

days, many serious games or GWAP "Game With A Purpose" (von Ahn and Dabbish, 2008) projects exist in different domains, like Open Mind Common Sense (Singh et al., 2002), ESP games, Learner (Chklovski and Gil, 2005), or CYC project [1]. Concerning more specifically lexical networks, similar projects exist like "small world of words"[2] launched in 2003 by KU Leuwen. For the moment, this project is limited to building relations of only one kind: associated ideas.

Looking at the Wikipedia project, the idea of building lexical resources with the help voluntary contributors comes to mind. Unfortunately, the Wikipedia paradigm cannot be easily applied to build a dictionary with rich lexical information. In Wikipedia, articles do not need to follow the same structure, while in a dictionary, the same structure and linguistic theory must be applied to all the articles. Moreover, while it is easy to contribute to an encyclopedia entry, not everyone has the linguistic knowledge to contribute to a dictionary. On reading Wiktionary entries, one realizes that the quality cannot be compared to existing paper dictionaries.

When looking at people playing online games through the Internet, one could think that it would be interesting to use this time for playing a game that would build lexical data in the background, specifically data that is difficult to find in existing dictionaries. In this context, the idea of a serious lexical game emerged. The first version was launched for French in 2007 (Lafourcade and Joubert, 2008), which has now around 250,000 nodes and 1,330,000 relations.

Our game aims at building a rich and evolving lexical network comparable to the famous English wordnet (Miller et al., 1990). The principle is as follows: a player *A* initiates a game, an instruction is displayed concerning a type of competency corresponding to a lexical relation (e.g. synonym, antonym, domain, intensifier) and a word *W* is chosen randomly in the database. Player *A* has then a limited amount of time for giving propositions that answer the instruction applied to the word *W*.

The same word *W* with the same instruction is proposed to another player *B* and the process is the same. The two half-games of player *A* and player *B* are asynchronous. For each common answer in *A* and *B*'s propositions, the two players earn a certain amount of points and credits. For the word *W*, the common answers of *A* and *B* players are entered into the database. This process participates in the construction of a lexical network linking terms with typed and weighted relations, validated by pairs of players. The relations are typed by the instructions given to the players and weighted with the number of pair players that proposed them. A more detailed description of the game in French is provided by Lafourcade and Zampa (2009).

3 Portuguese Version

The game interface was translated by a native Portuguese speaker. A preliminary step was to internationalize the text messages by separating them from the interface and storing them in an array, allowing for easy translation in any other language. Simultaneously, we developed, and tested an easy step-by-step installer which makes the deployment of the game as easy as installing a content management system software on a server.

A list of seed words must be provided from which the game will chose the proposed terms at the beginning. As the game evolves, people suggest new words not necessarily in the initial dictionary, thus helping the vocabulary to grow. Two resources were used to compose this list of seed words. The first is the DELAS–PB dictionary from NILC (Muniz, 2004). All nouns, verbs, adjectives and adverbs were extracted, resulting in 67,062 words. As these include a large number of rare words, pilot tests showed that the game became annoying when the player ignored the meaning of most of the proposed words. Therefore, the number of Google hits for every word was obtained and only the 20% most common ones were kept, resulting in a list of 13,413 words. To this, the entries of the Brazilian Open Mind Common Sense network (Anacleto et al., 2008) were added, in order to allow future comparison with this resource. Apertium's lt-toolbox[3] was used in order to obtain the most frequent POS tag for each entry, resulting in 5,129 nouns, 3,672 verbs, 1,176 adjectives, and 201 adverbs. The union with the preceding dictionary resulted in a final seed

[1] http://game.cyc.com/
[2] http://www.smallworldofwords.com/
[3] http://wiki.apertium.org/wiki/Lttoolbox

Collaboratively Building Language Resources while Localising the Web

Asanka Wasala, Reinhard Schäler, Ruvan Weerasinghe* and Chris Exton
Centre for Next Generation Localisation/Localisation Research Centre
CSIS Department, University of Limerick, Limerick, Ireland
*University of Colombo School of Computing, 35, Reid Avenue, Colombo 00700, Sri Lanka
{Asanka.Wasala, Reinhard.Schaler, Chris.Exton}@ul.ie,
*arw@ucsc.cmb.ac.lk

Abstract

In this paper, we propose the collaborative construction of language resources (translation memories) using a novel browser extension-based client-server architecture that allows translation (or 'localisation') of web content capturing and aligning source and target content produced by the 'power of the crowd'. The architectural approach chosen enables collaborative, in-context, and real-time localisation of web content supported by the crowd and high-quality language resources. To the best of our knowledge, this is the only practical web content localisation methodology currently being proposed that incorporates the collaborative construction and use of TMs. The approach also supports the building of resources such as parallel corpora – resources that are still not available for many, and especially not for underserved languages.

1 Introduction

A vast amount of knowledge is available on the web, primarily in English. There are millions of people worldwide, who cannot assimilate this knowledge mainly due the language service barrier. Although English is still dominating the web, the situation is changing. Non-English content is growing rapidly (Large and Moukdad, 2000; Daniel Brandon, 2001; Wasala and Weerasinghe, 2008).

Localisation is the translation and adaptation of digital content. Localisation of a website involves "translating text, content and adjusting graphical and visual elements, content and examples to make them culturally appropriate" (Stengers et al., 2004). However, the scope of our research is limited to the translation of text, which is arguably the most crucial component of web content localisation.

The study of web content localisation is a relatively new field within academia (Jiménez-Crespo, 2011). The only reported approaches to website localisation are human (Daniel Brandon, 2001) and machine-based translation (Large and Moukdad, 2000; Daniel Brandon, 2001; Wasala and Weerasinghe, 2008), with only very basic collaborative (Horvat, 2012) or first in-context approaches (Boxma, 2012) attempted. Although researchers have reported on the use of Machine Translation (MT) in web content localisation (Gaspari, 2007), the low quality of the MT-based website translation solutions is known to have been a significant drawback (Large and Moukdad, 2000; Daniel Brandon, 2001). Moreover, the research and development of MT systems for less-resourced languages is still in its infancy (Wasala and Weerasinghe, 2008). Therefore, MT-based web content localisation solutions are clearly not viable for less-resourced languages.

Undoubtedly, Web 2.0 and the constant increase of User Generated Content (UGC) lead to a higher demand for translation. The trend of crowdsourcing/social translation came into play only in the last few years. In this paper, we focus on crowdsourcing translation, i.e. when the crowd or a motivated part of it, participates in an open call to translate some content, creating highly valuable language resources in the process.

Browser extensions enhance the functionality of web browsers. Various browser extensions already exist that are capable of utilising existing Machine Translation (MT) services to translate web content into different languages. We exploit the power of

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 15–19,
Jeju, Republic of Korea, 8-14 July 2012.

browser extensions to design a conceptual localization layer for the web. Our research is mainly inspired by the works of Exton et al. (2009) on real-time localisation of desktop software using the crowd, Wasala and Weerasinghe (2008) on browser based pop-up dictionary extension, and Schäler on information sharing across languages (2012a) as well as social localisation (2012b).

The proposed architecture enables in-context real-time localisation of web content by communities sharing not just their content but also their language skills. The ultimate aim of this work is the collaborative creation of TMs which will allow for the automatic translation of web content based on reviewed and quality-checked, human produced translations. To the best of the authors' knowledge, this is the first effort of its kind to utilise the power of browser extensions along with TMs to build a website independent conceptual localisation layer with the aid of crowdsourcing.

The rest of the paper is organized as follows: Section 2 describes the architecture of the proposed system in detail; the development of the prototype is discussed in section 3; section 4 discusses key outstanding challenges and constraints of the proposed architecture; and finally, this paper concludes with a summary and discussion of future research directions.

2 System Architecture

In this section, the main functionalities of the proposed system architecture are described in detail.

The proposed system architecture is based on earlier work by Exton et al. (2009). They proposed a client-server architecture known as Update-Log-Daemon (UpLoD) for the localisation of applications' User Interface (UI) by the crowd. However, in our architecture, clients (browsers) connect to the central server via a browser extension. The browser extension implements the UpLoD architecture, which acts as a proxy between the browser and the central server.

We also extend the functionality of the central server in this architecture by equipping it with a component to maintain TMs for different language pairs.

2.1 Content Retrieval and Rendering Process

When the browser extension is installed and enabled, it allows a user to select the preferred locale. When a new URL is typed in, the browser will download the page. As soon as the content is downloaded, the browser extension will consult the central server for any TM matches in the user's preferred locale for the relevant URL. The TM matches will be retrieved with the contextual information. The next step is to replace the original content with the retrieved TM matches. With the aid of contextual hints that it received, the TM matches (i.e. target strings) will be replaced with the source strings. Finally, the content will be rendered in the browser. The contextual information may include: URL, last update date/time stamp, surrounding text with and without tags, XPath location of the segment, CSS properties among others as this information will helpful to precisely locate HTML elements in a web page (Selenium 2012). For replacing the original text with target strings, techniques such as Regular-expressions matching and XPath queries may be utilized.

2.2 Content Translation Process

The browser extension also facilitates the in-context translation of source content. Right clicking on a selected text will bring up a contextual menu where a "Translate" sub-menu can be found.

The extension allows in-context translation of the selected content segment in an editing environment similar to Wikipedia. Once the translation is completed, the extension sends the translated segment, original content and contextual information including URL to the central sever. Upon receiving translations from a client, the central server stores all the information that it retrieves in a TM.

The central server can be scheduled to periodically leverage translations as the TMs grow. Furthermore, later on, MT systems can be trained from the TM data and these trained MT systems can feed back into the system to speed up the translation process as well as to translate the content where TM matches are not found.

2.3 Translation Editing and Voting Process

As in the case of software localisation (Exton et al., 2009), a mechanism has to be built to choose the most appropriate translation of a given text segment. To assist in selecting the best translation for a given segment, a voting mechanism is proposed.

Resolving Task Specification and Path Inconsistency in Taxonomy Construction

Hui Yang
Department of Computer Science
Georgetown University
37th and O street NW
Washington, DC, 20057
huiyang@cs.georgetown.edu

Abstract

Taxonomies, such as Library of Congress Subject Headings and Open Directory Project, are widely used to support browsing-style information access in document collections. We call them browsing taxonomies. Most existing browsing taxonomies are manually constructed thus they could not easily adapt to arbitrary document collections. In this paper, we investigate both automatic and interactive techniques to derive taxonomies from scratch for arbitrary document collections. Particular, we focus on encoding user feedback in taxonomy construction process to handle task-specification rising from a given document collection. We also addresses the problem of path inconsistency due to local relation recognition in existing taxonomy construction algorithms. The user studies strongly suggest that the proposed approach successfully resolve task specification and path inconsistency in taxonomy construction.

1 Introduction

Taxonomies, such as Library of Congress Subject Headings (LCSH, 2011) and Open Directory Project (ODP, 2011), are widely used to support browsing-style information access in document collections. We call them browsing taxonomies. Browsing taxonomies are tree-structured hierarchies built upon a given document collection. Each term in a browsing hierarchy categorizes a set of documents related to this term. Driven by their needs, users can navigate through a the hierarchical structure of a browsing taxonomy to access particular documents. A browsing taxonomy can benefit information access via (1) providing an overview of (important) concepts in a document collection, (2) increasing the visibility of documents ranked low in a list (e.g. documents ordered by search relevance), and (3) presenting together documents about the same concept to allow more focused reading.

Most existing browsing taxonomies are manually constructed thus they could not easily adapt to arbitrary document collections. However, it is not uncommon that document collections are given ad-hoc for specific tasks, such as search result organization in for individual search queries (Carpineto et al., 2009) and literature investigation for a new research topic (Chau et al., 2011). There is a necessity to explore automatic or interactive techniques to support *quick* construction of browsing taxonomies for arbitrary document collections.

Most research on automatic taxonomy construction focuses on identifying local relations between concept pairs (Etzioni et al., 2005; Pantel and Pennacchiotti, 2006). The infamous problem of *path inconsistency*, which are usually caused by the local nature of most relation recognition algorithms when building a taxonomy, commonly exists in current research. Oftentimes, when a connecting concept for two pairs of parent-child concepts has multiple senses or represent mixed perspectives, the problem shows up. For example, while *financial institute→bank* and *bank→river bank* are correct; the path *financial institute→bank→river bank* is semantically inconsistent.

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 20–28,
Jeju, Republic of Korea, 8-14 July 2012.

In this paper, we propose a semi-supervised distance learning method to construct task-specific taxonomies. Assuming that a user is present to construct a taxonomy for browsing, the proposed approach directly learns semantic distances from the manual guidance provided by the user to predict semantically meaningful browsing taxonomies. Moreover, We tackle path inconsistency by posing constraints over root-to-leaf paths in a hierarchy to ensure concept consistency within paths

The contributions of our work include:

- It offers an opportunity for handling task specifications.
- Unlike most algorithms, our work takes care of path consistency during taxonomy construction.

The remainder of this paper is organized as follows: Section 2 describes the related work. Section 3 details the proposed automated algorithm for taxonomy construction. Section 4 presents the interactive algorithm to incorporate user feedback under a supervised semantic distance learning framework. Section 5 describes the evaluation and Section 6 concludes the paper.

2 Related Work

Most research conducted in the NLP community focuses on extracting local relations between concept pairs (Hearst, 1992; Berland and Charniak, 1999; Ravichandran and Hovy, 2002; Girju et al., 2003; Etzioni et al., 2005; Pantel and Pennacchiotti, 2006; Kozareva et al., 2008). More recently, more attention has been paid in building full taxonomies. For example, (Kozareva and Hovy, 2010) proposed to connect local concept pairs by finding the longest path in a subsumption graph. Both (Snow et al., 2006) and (Yang and Callan, 2009) incrementally grew taxonomies by adding new concepts at optimal positions within the existing structures. Specifically, Snow et al. estimated conditional probabilities by using syntactic parse features and decided taxonomic structure via maximizing overall likelihood of taxonomy. Yang and Callan proposed the ME framework to model the semantic distance $d(c_x, c_y)$ between concepts c_x and c_y as a weighted combination of numerous lexical and semantic features: $\sum_j \text{weight}_j * \text{feature}_j(c_x, c_y)$ and determine the taxonomic structure by minimizing overall distances.

An advantage in ME is that it allows manipulations to concept positions by incorporating various constraints to taxonomic structures. For example, ME handled concept generality-specificity by learning different distance functions for general concepts (located at upper levels) and specific concepts (located at lower levels) in a taxonomy.

In the Information Retrieval (IR) community, browsing taxonomies. also often called browsing hierarchies or Web directories, has been studied as an alternative to the ranked list representation for search results by the Information Retrieval (IR) community. The proposed forms of browsing structures include topic clusters (Cutting et al., 1992) and monothetic concept hierarchies (Sanderson and Croft, 1999; Lawrie et al., 2001; Kummamuru et al., 2004; Carpineto et al., 2009). The latter uses single concepts to represent documents containing them and organizes the concepts into hierarchies; they are in fact taxonomies. The major drawback of these approaches is that they often fail to produce meaningful taxonomic structures due to neglecting the semantics among concepts. For instance, (Sanderson and Croft, 1999) used document frequency and (Lawrie et al., 2001) used conditional probability to derive *is-a* relations. Moreover, they also suffer from path inconsistency when building full taxonomies.

3 Browsing Taxonomy Construction

To build browsing taxonomy for a document collection, the first step is to extract the concepts. We take a simple but effective approach. We exhaustively examine the collection and output a large set of terms, formed by nouns, noun phrases, and named entities occurring >5 times in the collection. We then filter out invalid terms due to part-of-speech errors or misspelling by removing terms that occur <4 times out of the top 10 returned snippets when submitting the term to *google.com* as a search query. We further conflate similar terms into clusters using LSA (Bellegarda et al., 1996) and select the most frequent terms as concepts from each term group. We select the N most frequent concepts to form the concept set C. N usually ranges from 30 to 100. We assume that C contains all concepts in the browsing taxonomy;

EAGER: Extending Automatically Gazetteers for Entity Recognition

10mm
20mm
-10 mm 10 mm
Omer Gunes, Christian Schallhart, Tim Furche
Department of Computer Science,
Oxford University, Oxford OX1 3QD
firstname.lastname@cs.ox.ac.uk

Jens Lehmann, Axel Ngonga
Institute of Computer Science,
University of Leipzig, 04103 Leipzig
lastname@informatik.uni-leipzig.de

Abstract

Key to named entity recognition, the manual gazetteering of entity lists is a costly, error-prone process that often yields results that are incomplete and suffer from sampling bias. Exploiting current sources of structured information, we propose a novel method for extending minimal seed lists into complete gazetteers. Like previous approaches, we value WIKIPEDIA as a huge, well-curated, and relatively unbiased source of entities. However, in contrast to previous work, we exploit not only its content, but also its structure, as exposed in DBPEDIA. We extend gazetteers through Wikipedia categories, carefully limiting the impact of noisy categorizations. The resulting gazetteers easily outperform previous approaches on named entity recognition.

1 Introduction

Automatically learning gazetteers with minimal supervision is a long standing problem in named entity recognition.

We propose EAGER as a novel approach to extending automatically gazetteers for entity recognition, utilizing DBPEDIA (Bizer et al., 2009) rather than WIKIPEDIA. DBPEDIA serves as a much better foundation than WIKIPEDIA, because all the information used in previous approaches (and much more) is already provided as a structured database of facts and articles. The extraction is more robust and complete than ad-hoc methods and maintained by a large community. E.g., navigating the category hierarchy is much easier and reliable with DBPEDIA.

To summarize, EAGER's main contributions are

(1) A *novel gazetteer expansion algorithm* that adds new entities from DBPEDIA. EAGER adds entities that have several categories in common with the seed terms, addressing noisy categorizations through a sophisticated *category pruning technique.*

(2) EAGER also extracts categories from DBPEDIA abstracts using *dependency analysis*. Finally, EAGER extracts plural forms and synonyms from *redirect information*.

(3) For *entity recognition*, we integrate the gazetteer with a simple, but effective machine learning classifier, and *experimentally* show that the extended gazetteers improve the F_1 score between 7% and 12% over our baseline approach and outperform (Zhang and Iria, 2009) on all learned concepts (subject, location, temporal).

2 Related Work

We divide the related work in automatic gazetteer population into three groups: **(1)** *Machine learning* approaches **(2)** *Pattern driven* approaches Finally, like our own work, **(3)** *knowledge driven* approaches

Knowledge Driven. In any case, machine learning and pattern driven approaches extract their terms from unstructured sources – despite the fact that large, general knowledge bases became available in the last years. One of the first knowledge-driven methods (Magnini et al., 2002) employed WORDNET to identify trigger words and candidate

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 29–33,
Jeju, Republic of Korea, 8-14 July 2012.

gazetteer terms with its word-class and -instance relations. As WORDNET covers domain specific vocabularies only to a limited extent, this approach is also limited in its general applicability.

In (Toral and Muñoz, 2006), gazetteers are built from the noun phrases in the first sentences of WIKIPEDIA articles by mapping these phrases to WORDNET and adding further terms found along the hypernymy relations. The approach presented in (Kazama and Torisawa, 2007; Kazama and Torisawa, 2008) relies solely on WIKIPEDIA, producing gazetteers without explicitly named concepts, arguing that consistent but anonymous labels are still useful.

Most closely related to our own work, the authors of (Zhang and Iria, 2009) build an approach solely on WIKIPEDIA which does not only exploit the article text but also analyzes the structural elements of WIKIPEDIA:

3 Automatically Extending Gazetteer Lists

3.1 Extraction Algorithm: Overview

Algorithm 1 shows an outline of the gazetteer expansion algorithm used in EAGER. To extend an initial seed set $\mathscr{S}$ EAGER proceeds, roughly, in three steps: First, it identifies DBPEDIA articles for seed entities and extracts implicit category and synonym information from abstracts and redirect information (Lines 1–11). Second, it finds additional categories from the DBPEDIA category hierarchy (Lines 12–20). Finally, it uses the categories from the first two steps to extract additional entities (Lines 21–24). In the following, we consider the three steps separately.

3.2 Implicit: Abstract and Redirects

Before EAGER can analyse abstract and redirect information for an article, we need to **find the corresponding DBPEDIA articles** (Lines 1–3) for each seed entry in $\mathscr{S}$. There may be one or more such entry. Here, we observe the first advantage of DBPEDIA's more structured information: DBPEDIA already contains plain text labels such as "Barack Obama" and we can directly query (using the SPARQL endpoint) all articles with a label equal (or starting with) an entity in our seed set. This allows for more precise article matching and avoids complex URL encodings as necessary in previous,

Algorithm 1: $\texttt{GazetteerExtension}(\mathscr{S})$

```
1  foreach seed entity e ∈ 𝒮 do
2      find article a for e in DBPEDIA;
3      Articles(e) ← a;
4  𝒢 ← ∅; 𝒫 ← ∅;
5  foreach entity e, article a = Articles(e) do
6      foreach sentence s ∈ a.Abstract do
7          D_s ← dependencies in s;
8          add all t : nsubj(e,t) ∈ D_s to 𝒫;
9          add all t : nsubj(e,t'), conj(t',t) ∈ D_s to 𝒫;
10     foreach article a' ∈ a.Redirects do
11         add all labels of a to 𝒢;
12     Cats(e) ← Cats(e) ∪ a.Cats;
13 foreach entity e, category c ∈ Cats(e) do
14     Cats(e) ← Cats(e) ∪ CategoryNeighbors(c,k);
15 foreach category c ∈ Cats(e) for some e do
16     Support(c) ← |{e' : c ∈ Cats(e')}|;
17 foreach connected component 𝒞 in Cats do
18     Support(𝒞) ← Σ_{c∈𝒞} Support(c);
19 MaxCatComp ← 𝒞 with maximal Support;
20 add all categories in MaxCatComp to 𝒫;
21 foreach category c ∈ 𝒫 do
22     foreach article a with c ∈ a.Cats do
23         if |a.Cats \ 𝒫| ≤ θ then
24             add all labels of a to 𝒢;
```

WIKIPEDIA-based approaches such as (Kazama and Torisawa, 2007). As (Zhang and Iria, 2009), we reject redirection entries in this step as ambiguous.

With the articles identified, we can proceed to extract category information from the abstracts and new entities from the redirect information. In the **dependency analysis of article abstracts** (Lines 6–9), we aim to extract category (or, more generally, hypernym) information from the abstracts of articles on the ssed list. We perform a standard dependency analysis on the sentences of the abstract and return all nouns that stand in `nsubj` relation to a seed entity or (directly or indirectly) in `conj` (correlative conjunction) relation to a noun that stands in `nsubj` relation to a seed entity. This allows us to extract, e.g., both "general" and "statesman" as categories from a sentence such as "Julius Caesar was a Roman general and statesman". This analysis is inspired by (Zhang and Iria, 2009), but performed on the entire abstract which is clearly dis-

Extracting Context-Rich Entailment Rules from Wikipedia Revision History

Elena Cabrio
INRIA
2004, route de Lucioles BP93
06902 Sophia Antipolis, France.
elena.cabrio@inria.fr

Bernardo Magnini
FBK
Via Sommarive 18
38100 Povo-Trento, Italy.
magnini@fbk.eu

Angelina Ivanova
University of Oslo
Gaustadalléen 23B
Ole-Johan Dahls hus
N-0373 Oslo, Norway.
angelii@ifi.uio.no

Abstract

Recent work on Textual Entailment has shown a crucial role of knowledge to support entailment inferences. However, it has also been demonstrated that currently available entailment rules are still far from being optimal. We propose a methodology for the automatic acquisition of large scale context-rich entailment rules from Wikipedia revisions, taking advantage of the syntactic structure of entailment pairs to define the more appropriate linguistic constraints for the rule to be successfully applicable. We report on rule acquisition experiments on Wikipedia, showing that it enables the creation of an innovative (i.e. acquired rules are not present in other available resources) and good quality rule repository.

1 Introduction

Entailment rules have been introduced to provide pieces of knowledge that may support entailment judgments (Dagan *et al.*, 2009) with some degree of confidence. More specifically, an entailment rule is defined (Szpektor *et al.*, 2007) as a directional relation between two sides of a pattern, corresponding to text fragments with variables (typically phrases or parse sub-trees). The left-hand side (LHS) of the pattern entails the right-hand side (RHS) of the same pattern under the same variable instantiation. Given the Text-Hypothesis pair (T-H) in Example 1:

Example 1.

T: *Dr. Thomas Bond established a hospital in Philadelphia for the reception and cure of poor sick persons.*

H: *Dr. Bond created a medical institution for sick people.*

a (directional) lexical rule like:

1) **LHS:** *hospital* ⇒ **RHS:** *medical institution*
probability: 0.8

brings to a TE system (aimed at recognizing that a particular target meaning can be inferred from different text variants in several NLP application, e.g. Question Answering or Information Extraction) the knowledge that the word *hospital* in Text can be aligned, or transformed, into the word *medical institution* in the Hypothesis, with a probability 0.8 that this operation preserves the entailment relation among T and H. Similar considerations apply for more complex rules involving verbs, as:

2) **LHS:** *X establish Y* ⇒ **RHS:** *X create Y*
probability: 0.8

where the variables may be instantiated by any textual element with a specified syntactic relation with the verb. Both kinds of rules are typically acquired either from structured sources (e.g. WordNet (Fellbaum, 1998)), or from unstructured sources according for instance to distributional properties (e.g. DIRT (Lin and Pantel, 2001)). Entailment rules should typically be applied only in specific contexts, defined in (Szpektor *et al.*, 2007) as *relevant contexts*. Some existing paraphrase and entailment acquisition algorithms add constraints to the learned rules (e.g. (Sekine, 2005), (Callison-Burch, 2008)), but most do not. Because of a lack of an adequate representation of the linguistic context in which the

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 34–43,
Jeju, Republic of Korea, 8-14 July 2012.

rules can be successfully applied, their concrete use reflects this limitation. For instance, rule 2 (extracted from DIRT) fails if applied to "The mathematician established the validity of the conjecture", where the sense of *establish* is not a synonym of *create* (but of *prove, demonstrate*), decreasing system's precision. Moreover, these rules often suffer from lack of directionality, and from low accuracy (i.e. the strength of association of the two sides of the rule is often weak, and not well defined). Such observations are also in line with the discussion on ablation tests carried out at the last RTE evaluation campaigns (Bentivogli *et al.*, 2010).

Additional constraints specifying the variable types are therefore required to correctly instantiate them. In this work, we propose to take advantage of Collaboratively Constructed Semantic Resources (CSRs) (namely, Wikipedia) to mine information useful to context-rich entailment rule acquisition. More specifically, we take advantage of material obtained through Wikipedia revisions, which provides at the same time real textual variations from which we may extrapolate the relevant syntactic context, and several simplifications with respect to alternative resources. We consider T-H pairs where T is a revision of a Wikipedia sentence and H is the original sentence, as the revision is considered more informative then the revised sentence.

We demonstrate the feasibility of the proposed approach for the acquisition of context-rich rules from Wikipedia revision pairs, focusing on two case studies, i.e. the acquisition of entailment rules for *causality* and for *temporal expressions*. Both phenomena are highly frequent in TE pairs, and for both there are no available resources yet. The result of our experiments consists in a repository that can be used by TE systems, and that can be easily extended to entailment rules for other phenomena.

The paper is organized as follows. Section 2 reports on previous work, highlighting the specificity of our work. Section 3 motivates and describes the general principles underlying our acquisition methodology. Section 4 describes in details the steps for context-rich rules acquisition from Wikipedia pairs. Section 5 reports about the experiments on causality and temporal expressions and the obtained results. Finally, Section 6 concludes the paper and suggests directions for future improvements.

2 Related work

The use of Wikipedia revision history in NLP tasks has been previously investigated by a few works. In (Zanzotto and Pennacchiotti, 2010), two versions of Wikipedia and semi-supervised machine learning methods are used to extract large TE data sets similar to the ones provided for the RTE challenges. (Yatskar *et al.*, 2010) focus on using edit histories in Simple English Wikipedia to extract lexical simplifications. Nelken and Yamangil (2008) compare different versions of the same document to collect users' editorial choices, for automated text correction, sentence compression and text summarization systems. (Max and Wisniewski, 2010) use the revision history of French Wikipedia to create a corpus of natural rewritings, including spelling corrections, reformulations, and other local text transformations. In (Dutrey *et al.*, 2011), a subpart of this corpus is analyzed to define a typology of local modifications.

Because of its high coverage, Wikipedia is used by the TE community for lexical-semantic rules acquisition, named entity recognition, geographical information[1] (e.g. (Mehdad *et al.*, 2009), (Mirkin *et al.*, 2009), (Iftene and Moruz, 2010)), i.e. to provide TE systems with world and background knowledge. However, so far it has only been used as source of factual knowledge, while in our work the focus is on the acquisition of more complex rules, concerning for instance spatial or temporal expressions.

The interest of the research community in producing specific methods to collect inference and paraphrase pairs is proven by a number of works in the field, which are relevant to the proposed approach.

As for paraphrase, Sekine's Paraphrase Database (Sekine, 2005) is collected using an unsupervised method, and focuses on phrases connecting two Named Entities. In the Microsoft Research Paraphrase Corpus[2], pairs of sentences are extracted from news sources on the web, and manually annotated. As for rule repositories collected using distributional properties, DIRT (Discovery of Inference Rules from Text)[3] is a collection of inference rules

[1] http://www.aclweb.org/aclwiki/index.php?title=RTE_Knowledge_Resources

[2] http://research.microsoft.com/en-us/downloads

[3] http://www.aclweb.org/aclwiki/index.php?title=DIRT_Paraphrase_Collection