ACL-IJCNLP 2009

Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP

Proceedings of Software Demonstrations

3 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

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

Preface

Welcome to the proceedings of the demo session. This volume contains the abstracts of the software demonstrations presented at the combined 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing, held in Suntec, Singapore, on August 3.

The demonstrations program offers the presentation of early research prototypes as well as interesting mature systems. The demo chairs and the members of the program committee received 23 submissions, 11 of which were selected for inclusion in the program after review by at least two members of the program committee and on-line discussions for nine boundary papers.

We would like to thank the members of our program committee for their excellent job in reviewing the submissions and providing their support in the decisive discussions.

Co-chairs:

Gary Geunbae Lee (Pohang University of Science and Technology, South Korea)
Sabine Schulte im Walde (University of Stuttgart, Germany)

Program Committee:

Paul Buitelaar (DERI, Ireland)
Massimiliano Ciaramita (Google, Switzerland)
Sadao Kurohasi (Kyoto University, Japan)
Ee-Peng Lim (Singapore Management University, Singapore)
Dekang Lin (Google, USA)
Jong Park (KAIST, Korea)
Ted Pedersen (University of Minnesota, USA)
Dan Roth (University of Illinois at Urbana-Champaign, USA)
Ming Zhou (MSRA, China)
Heike Zinsmeister (University of Konstanz, Germany)

Table of Contents

WISDOM: A Web Information Credibility Analysis System

Susumu Akamine[†] **Daisuke Kawahara**[†] **Yoshikiyo Kato**[†]
Tetsuji Nakagawa[†] **Kentaro Inui**[†] **Sadao Kurohashi**[†‡] **Yutaka Kidawara**[†]

[†]National Institute of Information and Communications Technology
[‡]Graduate School of Informatics, Kyoto University
{akamine, dk, ykato, tnaka, inui, kidawara}@nict.go.jp, kuro@i.kyoto-u.ac.jp

Abstract

We demonstrate an information credibility analysis system called WISDOM. The purpose of WISDOM is to evaluate the credibility of information available on the Web from multiple viewpoints. WISDOM considers the following to be the source of information credibility: information contents, information senders, and information appearances. We aim at analyzing and organizing these measures on the basis of semantics-oriented natural language processing (NLP) techniques.

1. Introduction

As computers and computer networks become increasingly sophisticated, a vast amount of information and knowledge has been accumulated and circulated on the Web. They provide people with options regarding their daily lives and are starting to have a strong influence on governmental policies and business management. However, a crucial problem is that the information available on the Web is not necessarily credible. It is actually very difficult for human beings to judge the credibility of the information and even more difficult for computers. However, computers can be used to develop a system that collects, organizes, and relativises information and helps human beings view information from several viewpoints and judge the credibility of the information.

Information organization is a promising endeavor in the area of next-generation Web search. The search engine Clusty provides a search result clustering[1], and Cuil classifies a search result on the basis of query-related terms[2]. The persuasive technology research project at Stanford University discussed how websites can be designed to influence people’s perceptions (B. J. Fogg, 2003). However, as per our knowledge, no research has been carried out for supporting the human judgment on information credibility and information organization systems for this purpose.

In order to support the judgment of information credibility, it is necessary to extract the background, facts, and various opinions and their distribution for a given topic. For this purpose, syntactic and discourse structures must be analyzed, their types and relations must be extracted, and synonymous and ambiguous expressions should be handled properly.

Furthermore, it is important to determine the identity of the information sender and his/her specialty as criteria for credibility, which require named entity recognition and total analysis of documents.

In this paper, we describe an information credibility analysis system called WISDOM, which automatically analyzes and organizes the above aspects on the basis of semantically oriented NLP techniques. WISDOM currently operates over 100 million Japanese Web pages.

2. Overview of WISDOM

We consider the following three criteria for the judgment of information credibility.

(1) Credibility of information contents,
(2) Credibility of the information sender, and
(3) Credibility estimated from the document style and superficial characteristics.

In order to help people judge the credibility of information from these viewpoints, we have been developing an information analysis system called WISDOM. Figure 1 shows the analysis result of WISDOM on the analysis topic “Is bio-ethanol good for the environment?” Figure 2 shows the system architecture of WISDOM.

Given an analysis topic (query), WISDOM sends the query to the search engine TSUBAKI (Shinzato et al., 2008), and TSUBAKI returns a list of the top N relevant Web pages (N is usually set to 1000).

Then, those pages are automatically analyzed, and major and contradictory expressions and evaluative expressions are extracted. Furthermore, the information senders of the Web pages, which were analyzed beforehand, are collected and the distribution is calculated.

The WISDOM analysis results can be viewed from several viewpoints by changing the tabs using a Web browser. The leftmost tab, “Summary,” shows the summary of the analysis, with major phrases and major/contradictory statements first.

[1] http://clusty.com/, http://clusty.jp/

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 1–4,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP

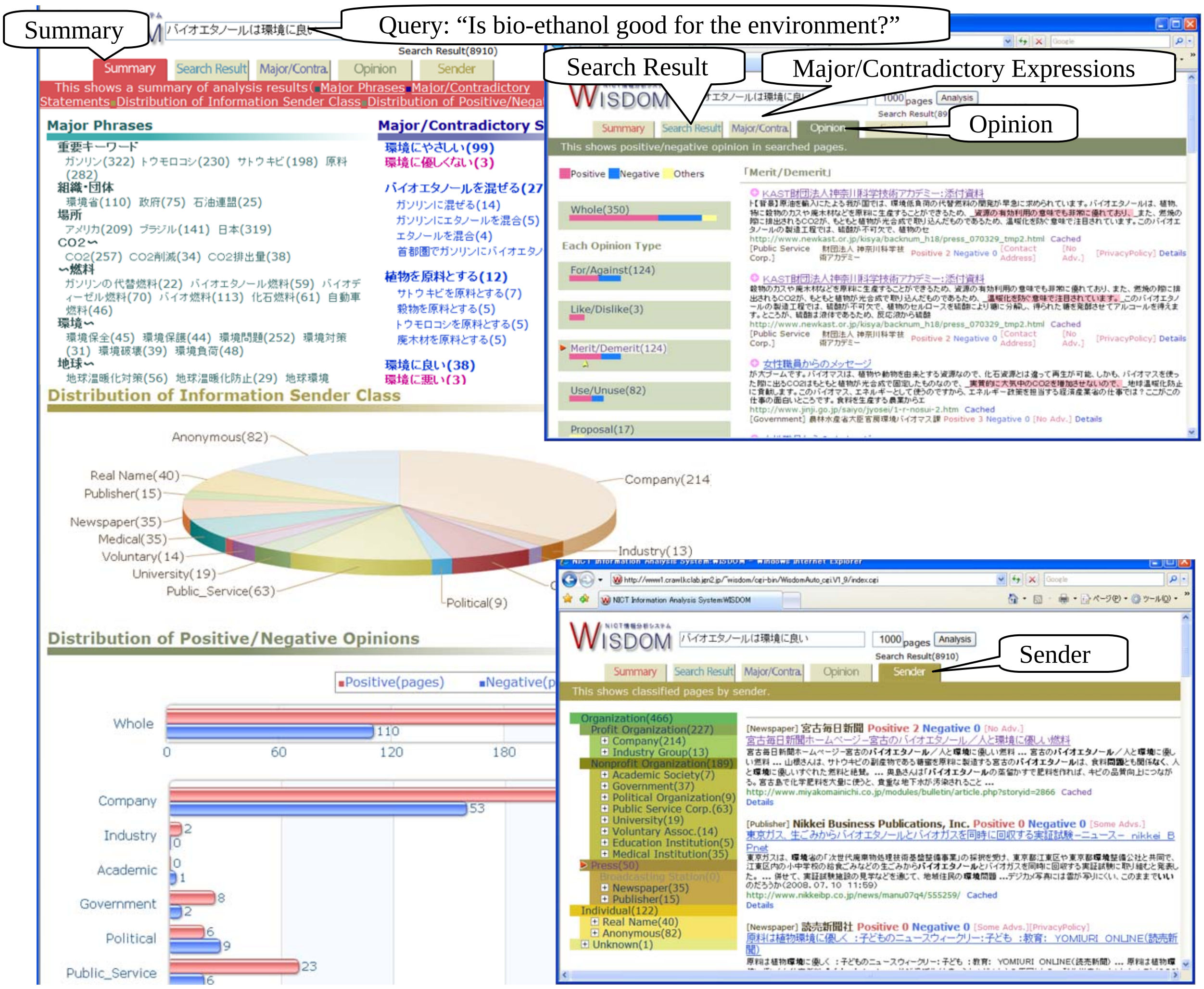

Summary
Query: “Is bio-ethanol good for the environment?”
Search Result
Major/Contradictory Expressions
Opinion
Sender
バイオエタノールは環境に良い
Search Result(8910)
Summary
Search Result
Major/Contra.
Opinion
Sender
This shows a summary of analysis results(Major Phrases Major/Contradictory Statements Distribution of Information Sender Class Distribution of Positive/Nega
Major Phrases
重要キーワード
ガソリン(322) トウモロコシ(230) サトウキビ(198) 原料(282)
組織・団体
環境省(110) 政府(75) 石油連盟(25)
場所
アメリカ(209) ブラジル(141) 日本(319)
CO2〜
CO2(257) CO2削減(34) CO2排出量(38)
〜燃料
ガソリンの代替燃料(22) バイオエタノール燃料(59) バイオディーゼル燃料(70) バイオ燃料(113) 化石燃料(61) 自動車燃料(46)
環境〜
環境保全(45) 環境保護(44) 環境問題(252) 環境対策(31) 環境破壊(39) 環境負荷(48)
地球〜
地球温暖化対策(56) 地球温暖化防止(29) 地球環境
Major/Contradictory S
環境にやさしい(99)
環境に優しくない(3)
バイオエタノールを混ぜる(27
ガソリンに混ぜる(14)
ガソリンにエタノールを混合(5)
エタノールを混合(4)
首都圏でガソリンにバイオエタノ
植物を原料とする(12)
サトウキビを原料とする(7)
穀物を原料とする(5)
トウモロコシを原料とする(5)
廃木材を原料とする(5)
環境に良い(38)
環境に悪い(3)
Distribution of Information Sender Class
Anonymous(82)
Real Name(40)
Publisher(15)
Newspaper(35)
Medical(35)
Voluntary(14)
University(19)
Public_Service(63)
Company(214
Industry(13)
Political(9)
Distribution of Positive/Negative Opinions
Positive(pages)
Negative(p
Whole
110
0
60
120
180
Company
53
Industry
2
0
Academic
0
1
Government
8
2
Political
6
9
Public_Service
23
6
WISDOM
This shows positive/negative opinion in searched pages.
Positive
Negative
Others
Whole(350)
Each Opinion Type
For/Against(124)
Like/Dislike(3)
Merit/Demerit(124)
Use/Unuse(82)
Proposal(17)
「Merit/Demerit」
1000 pages
Analysis
This shows classified pages by sender.
Organization(466)
Profit Organization(227)
Company(214)
Industry Group(13)
Nonprofit Organization(189)
Academic Society(7)
Government(37)
Political Organization(9)
Public Service Corp.(63)
University(19)
Voluntary Assoc.(14)
Education Institution(5)
Medical Institution(35)
Press(50)
Broadcasting Station(0)
Newspaper(35)
Publisher(15)
Individual(122)
Real Name(40)
Anonymous(82)
Unknown(1)
[Newspaper] 宮古毎日新聞 Positive 2 Negative 0
[Publisher] Nikkei Business Publications, Inc. Positive 0 Negative 0
[Newspaper] 読売新聞社 Positive 0 Negative 0

Figure 1. An analysis example of the information credibility analysis system WISDOM.

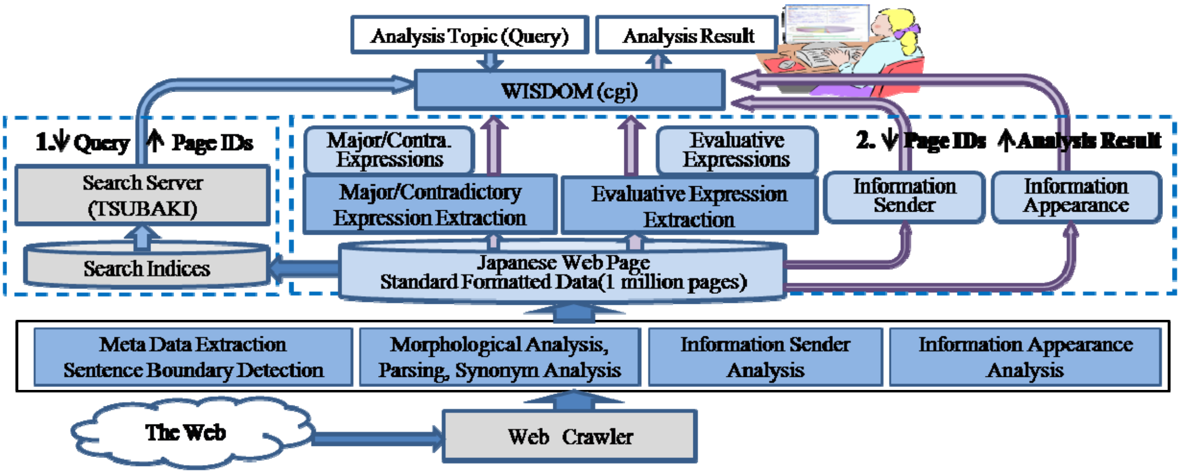

Analysis Topic (Query)
Analysis Result
WISDOM (cgi)
1. Query
Page IDs
Search Server (TSUBAKI)
Search Indices
Major/Contra. Expressions
Major/Contradictory Expression Extraction
Evaluative Expressions
Evaluative Expression Extraction
2. Page IDs
Analysis Result
Information Sender
Information Appearance
Japanese Web Page Standard Formatted Data(1 million pages)
Meta Data Extraction Sentence Boundary Detection
Morphological Analysis, Parsing, Synonym Analysis
Information Sender Analysis
Information Appearance Analysis
The Web
Web Crawler

Figure 2. System architecture of WISDOM.

By referring to these phrases and statements, a user can grasp the important issues related to the topic at a glance. The pie diagram indicates the distribution of the information sender class spread over 1000 pages, such as company, industry group, and government. The names of the information senders of the class can be viewed by placing the cursor over a class region. The last bar chart shows the distribution of positive and negative opinions related to the topic spread over 1000 pages, for all and for each sender class. For example, with regard to “Bio-ethanol,” we can see that the number of positive opinions is more than that of negative opinions, but it is the opposite in the case of some sender classes. Several display units in the Summary tab are cursor sensitive, providing links to more detailed information (e.g., the page list including a major statement, the page list of a sender class, and the page list containing negative opinions).

The “Search Result” tab shows the search result by TSUBAKI, i.e., ranking the relevant pages according to the TSUBAKI criteria. The “Major/Contradictory Expressions” tab shows the list of major phrases and major/contradictory statements about the given topic and the list of pages containing the specified phrase or statement. The “Opinion” tab shows the analysis result of the evaluative expressions, classified according to for/against, like/dislike, merit/demerit, and others, and it also shows the list of pages containing the specified type of evaluative expressions. The “Sender” tab classifies the pages according to the class of the information sender, for example, a user can view the pages created only by the government.

Furthermore, the superficial characteristics of pages called as information appearance are analyzed beforehand and can be viewed in WISDOM, such as whether or not the contact address is shown in the page and the privacy policy is on the page, the volume of advertisements on the page, the number of images, and the number of in/out links.

As shown thus far, given an analysis topic, WISDOM collects and organizes the relevant information available on the Web and provides users with multi-faceted views. We believe that such a system can considerably support the human judgment of information credibility.

3. Data Infrastructure

We usually utilize 100 million Japanese Web pages as the analysis target. The Web pages have been converted into the standard formatted Web data, an XML format. The format includes several metadata such as URLs, crawl dates, titles, and in/out links. A text in a page is automatically segmented into sentences (note that the sentence boundary is not clear in the original HTML file), and the analysis results obtained by a morphological analyzer, parser, and synonym analyzer are also stored in the standard format. Furthermore, the site operator, the page author, and information appearance (e.g., contact address, privacy policy, volume of advertisements, and images) are automatically analyzed and stored in the standard format.

4. Extraction of Major Expressions and Their Contradictions

For the organization of information contents, WISDOM extracts and presents the major expressions and their contradictions on a given analysis topic (Kawahara et al., 2008). Major expressions are defined as expressions occurring at a high frequency in the set of Web pages on the analysis topic. They are classified into two: noun phrases and predicate-argument structures (statements). Contradictions are the predicate-argument structures that contradict the major expressions. For the Japanese phrase *yutori kyouiku* (cram-free education), for example, *tsumekomi kyouiku* (cramming education) and *ikiru chikara* (life skills) are extracted as the major noun phrases; *yutori kyouiku-wo minaosu* (reexamine cram-free education) and *gakuryokuga teika-suru* (scholastic ability deteriorates), as the major predicate-argument structures; and *gakuryoku-ga koujousuru* (scholastic ability ameliorates), as its contradiction. This kind of summarized information enables a user to grasp the facts and arguments on the analysis topic available on the Web.

We use 1000 Web pages for a topic retrieved from the search engine TSUBAKI. Our method of extracting major expressions and their contradictions consists of the following steps:

1. Extracting candidates of major expressions:

The candidates of major expressions are extracted from each Web page in the search result. From the relevant sentences to the analysis topic that consist of approximately 15 sentences selected from each Web page, compound nouns, parenthetical expressions, and predicate-argument structures are extracted as the candidates of the major expressions.

2. Distilling major expressions:

Simply presenting expressions at a high frequency is not always information of high quality. This is because scattering synonymous expressions such as *karikyuramu* (curriculum) and *kyouiku katei* (course of study) and entailing expressions such as IWC and IWC *soukai* (IWC plenary session), all of which occur frequently, hamper the understanding process of users. Further, synonymous predicate-argument structures such as *gakuryoku-ga teika-suru* (scholastic ability deteriorates) and *gakuryoku-ga sagaru* (scholastic ability lowers) have the same problem. To overcome this problem, we distill major expressions by merging spelling variations with morphological analysis, merging synonymous expressions automatically acquired from an ordinary dictionary and the Web, and merging expressions that can be entailed by another expression.

3. Extracting contradictory expressions:

Predicate-argument structures that negate the predicate of major ones and that replace the predicate of major ones with its antonym are extracted as contradictions. For example, *gakuryoku-ga teika-shi-nai* (scholastic ability does not deteriorate) and *gakuryokuga koujou-suru* (scholastic ability ameliorates) are extracted as the contradictions to *gakuryoku-ga teikasuru* (scholastic ability deteriorates). This process is performed using an antonym lexicon, which consists of approximately 2000 pairs; these pairs are extracted from an ordinary dictionary.

5. Extraction of Evaluative Information

The extraction and classification of evaluative information from texts are important tasks with many applications and they have been actively studied recently (Pang and Lee, 2008). Most previous studies on opinion extraction or sentiment analysis deal with only subjective and explicit expressions. For example, Japanese sentences such as *watashi-wa apple-ga sukida* (I like apples) and *kono seido-ni hantaida* (I oppose the system) contain evaluative expressions that are directly expressed with subjective expressions. However, sentences such as *kono shokuhin-wa kou-gan-kouka-ga aru* (this food has an anti-cancer effect) and *kono camera-wa katte 3-ka-de kowareta* (this camera was broken 3 days after I bought it) do not contain subjective expressions but contain negative evaluative expressions. From the viewpoint of information credibility, it appears important to deal with a wide variety of evaluative information including such implicit evaluative expressions (Nakagawa et al., 2008).

A corpus annotated with evaluative information was developed for evaluative information analysis studies. Fifty topics such as "Bio-ethanol" and "Pension plan" were chosen. For each topic, 200 sentences containing the topic word were collected from the Web to construct the corpus totaling 10,000 sentences. For each sentence, annotators judged whether or not the sentence contained evaluative expressions. When evaluative expressions were identified, the evaluative expressions, their holders, their sentiment polarities (positive or negative), and their relevance to the topic were annotated.

We developed an automatic analyzer of evaluative information using the corpus. We performed experiments of sentiment polarity classification using Support Vector Machines. Word forms, POS tags, and sentiment polarities from an evaluative word dictionary of all the words in evaluative expressions were used as features, and an accuracy of 83% was obtained. From the error analysis, we found that it was difficult to classify domain-specific evaluative expressions; we are now planning the automatic acquisition of evaluative word dictionaries.

6. Information Sender Analysis

The source of information (or information sender) is one of the important elements when judging the credibility of information. It is rather easy for human beings to identify the information sender of a Web page. When reading a Web page, whether it is deliberate or not, we attribute some characteristics to the information sender and accordingly form our attitudes toward the information. However, the state-of-the-art search engines do not provide facilities to organize a vast amount of information on the basis of the information sender. If we can organize the information on a topic on the basis of who or what type the information sender is, it would enable the user to grasp an overview of the topic or to judge the credibility of relevant information.

WISDOM automatically identifies the *site operators* of Web pages and classifies them into predefined categories of information sender called *information sender class*. A site operator of a Web page is the governing body of a website on which the page is published. The information sender class categorizes the information sender on the basis of axes such as individuals vs. organizations and profit vs. nonprofit organizations. The list below shows the categories of information sender class.

1. Organization
   (a) Profit Organization
      i. Company
      ii. Industry Group
   (b) Nonprofit Organization
      i. Academic Society
      ii. Government
      iii. Political Organization
      iv. Public Service Corp., Nonprofit Organization
      v. University
      vi. Voluntary Association
      vii. Education Institution

1. Organization (cont'd)
   (c) Press
      i. Broadcasting Station
      ii. Newspaper
      iii. Publisher
2. Individual
   (a) Real Name
   (b) Anonymous, Screen Name

WISDOM allows the user to organize the information on the basis of the information sender class assigned to each Web page. Technical details of the information sender analysis employed in WISDOM can be found in (Kato et al., 2008).

7. Conclusions

This paper has described an information analysis system called WISDOM. As shown in this paper, WISDOM already provides a reasonably nice organized view for a given topic and can serve as a useful tool for handling informational queries and for supporting human judgment of information credibility. WISDOM is freely available at http://wisdom-nict.jp/.

References

B. J. Fogg. 2003. *Persuasive Technology: Using Computers to Change What We Think and Do (The Morgan Kaufmann Series in Interactive Technologies).* Morgan Kaufmann.

K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi 2008. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Proceedings of IJCNLP2008.*

D. Kawahara, S. Kurohashi, and K. Inui 2008. Grasping major statements and their contradictions toward information credibility analysis of web contents. In *Proceedings of WI'08.*

B. Pang and L. Lee 2008. Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, Volume 2, Issue 1-2, 2008.

T. Nakagawa, T. Kawada, K. Inui, and S. Kurohashi 2008. Extracting subjective and objective evaluative expressions from the web. In *Proceedings of ISUC2008.*

Y. Kato, D. Kawahara, K. Inui, S. Kurohashi, and T. Shibata 2008. Extracting the author of web pages. In *Proceedings of WICOW2008.*

LX-Center: a center of online linguistic services

António Branco, Francisco Costa, Eduardo Ferreira, Pedro Martins,
Filipe Nunes, João Silva and Sara Silveira
University of Lisbon
Department of Informatics
{antonio.branco, fcosta, eferreira, pedro.martins,
fnunes, jsilva, sara.silveira}@di.fc.ul.pt

Abstract

This is a paper supporting the demonstration of the LX-Center at ACL-IJCNLP-09.

LX-Center is a web center of online linguistic services aimed at both demonstrating a range of language technology tools and at fostering the education, research and development in natural language science and technology.

1 Introduction

This paper is aimed at supporting the demonstration of a web center of online linguistic services. These services demonstrate language technology tools for the Portuguese language and are made available to foster the education, research and development in natural language science and technology.

This paper adheres to the common format defined for demo proposals: the next Section 2 presents an extended abstract of the technical content to be demonstrated; Section 3 provides a script outline of the demo presentation; and the last Section 4 describes the hardware and internet requirements expected to be provided by the local organizer.

2 Extended abstract

The LX-Center is a web center of online linguistic services for the Portuguese language located at `http://lxcenter.di.fc.ul.pt`. This is a freely available center targeted at human users. It has a counterpart in terms of a webservice for software agents, the LXService, presented elsewhere (Branco et al., 2008).

2.1 LX-Center

The LX-Center encompasses linguistic services that are being developed, in all or part, and maintained at the University of Lisbon, Department of Informatics, by the NLX-Natural Language and Speech Group. At present, it makes available the following functionalities:

- Sentence splitting
- Tokenization
- Nominal lemmatization
- Nominal morphological analysis
- Nominal inflection
- Verbal lemmatization
- Verbal morphological analysis
- Verbal conjugation
- POS-tagging
- Named entity recognition
- Annotated corpus concordancing
- Aligned wordnet browsing

These functionalities are provided by one or more of the seven online services that integrate the LX-Center. For instance, the LX-Suite service accepts raw text and returns it sentence splitted, tokenized, POS tagged, lemmatized and morphologically analyzed (for both verbs and nominals). Some other services, in turn, may support only one of the functionalities above. For instance, the LX-NER service ensures only named entity recognition.

These are the services offered by the LX-Center:

- LX-Conjugator
- LX-Lemmatizer
- LX-Inflector
- LX-Suite
- LX-NER
- CINTIL concordancer
- MWN.PT browser

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 5–8,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP

The access to each one of these services is obtained by clicking on the corresponding button on the left menu of the LX-Center front page.

Each of the seven services integrating the LX-Center will be briefly presented in a different subsection below. Fully fledged descriptions are available at the corresponding web pages and in the white papers possibly referred to there.

2.2 LX-Conjugator

The LX-Conjugator is an online service for fully-fledged conjugation of Portuguese verbs. It takes an infinitive verb form and delivers all the corresponding conjugated forms. This service is supported by a tool based on general string replacement rules for word endings supplemented by a list of overriding exceptions. It handles both known verbs and unknown verbs, thus conjugating neologisms (with orthographic infinitival suffix).

The Portuguese verbal inflection system is a most complex part of the Portuguese morphology, and of the Portuguese language, given the high number of conjugated forms for each verb (ca. 70 forms in non pronominal conjugation), the number of productive inflection rules involved and the number of non regular forms and exceptions to such rules.

This complexity is further increased when the so-called pronominal conjugation is taken into account. The Portuguese language has verbal clitics, which according to some authors are to be analyzed as integrating the inflectional suffix system: the forms of the clitics may depend on the Number (Singular vs. Plural), the Person (First, Second, Third or Second courtesy), the Gender (Masculine vs. Feminine), the grammatical function which they are in correspondence with (Subject, Direct object or Indirect object), and the anaphoric properties (Pronominal vs. Reflexive); up to three clitics (e.g. *deu-se-lho* / gave-One-ToHim-It) may be associated with a verb form; clitics may occur in so called enclisis, i.e. as a final part of the verb form (e.g. *deu-o* / gave-It), or in mesoclisis, i.e. as a medial part of the verb form (e.g. *dá-lo-ia* / give-it-Condicional) — when the verb form occurs in certain syntactic or semantic contexts (e.g in the scope of negation), the clitics appear in proclisis, i.e. before the verb form (ex.: *não o deu* / NOT it gave); clitics follow specific rules for their concatenation.

With LX-Conjugator, pronominal conjugation can be fully parameterizable and is thus exhaustively handled. Additionally, LX-Conjugator exhaustively handles a set of inflection cases which tend not to be supported together in verbal conjugators: Compound tenses; Double forms for past participles (regular and irregular); Past participle forms inflected for number and gender (with transitive and unaccusative verbs); Negative imperative forms; Courtesy forms for second person.

This service handles also the very few cases where there may be different forms in different variants: when a given verb has different orthographic representations for some of its inflected forms (e.g. *arguir* in European vs. *argüir* in American Portuguese), all such representations will be displayed.

2.3 LX-Lemmatizer

The LX-Lemmatizer is an online service for fully-fledged lemmatization and morphological analysis of Portuguese verbs. It takes a verb form and delivers all the possible corresponding lemmata (infinitive forms) together with inflectional feature values.

This service is supported by a tool based on general string replacement rules for word endings whose outcome is validated by the reverse procedure of conjugation of the output and matching with the original input. These rules are supplemented by a list of overriding exceptions. It thus handles an open set of verb forms provided these input forms bear an admissible verbal inflection ending. Hence, this service processes both lexically known and unknown verbs, thus coping with neologisms.

LX-Lemmatizer handles the same range of forms handled and generated by the LX-Conjugator. As for pronominal conjugation forms, the outcome displays the clitic detached from the lemma. The LX-Lemmatizer and the LX-Conjugator can be used in ”roll-over” mode. Once the outcome of say the LX-Conjugator on a given input lemma is displayed, the user can click over any one of the verbal forms in that conjugation table. This activates the LX-Lemmatizer on that input verb form, and then its possible lemmas, together with corresponding inflection feature values, are displayed. Now, any of these lemmas can also be clicked on, which will activate back the LX-Conjugator and will make the corresponding conjugation table to be displayed.

2.4 LX-Inflector

The LX-Inflector is an online service for the lemmatization and inflection of nouns and adjectives of Portuguese. This service is also based on a tool that relies on general rules for ending string replacement, supplemented by a list of overriding exceptions. Hence, it handles both lexically known and unknown forms, thus handling possible neologisms (with orthographic suffixes for nominal inflection).

As input, this service takes a Portuguese nominal form — a form of a noun or an adjective, including adjectival forms of past participles –, together with a bundle of inflectional feature values — values of inflectional features of Gender and Number intended for the output.

As output, it returns: inflectional features — the input form is echoed with the corresponding values for its inflectional features of Gender and Number, that resulted from its morphological analysis; lemmata — the lemmata (singular and masculine forms when available) possibly corresponding to the input form; inflected forms — the inflected forms (when available) of each lemma in accordance with the values for inflectional features entered. LX-Inflector processes both simple, prefixed or non prefixed, and compound forms.

2.5 LX-Suite

The LX-Suite is an online service for the shallow processing of Portuguese. It accepts raw text and returns it sentence splitted, tokenized, POS tagged, lemmatized and morphologically analyzed.

This service is based on a pipeline of a number of tools, including those supporting the services described above. Those tools, for lemmatization and morphological analysis, are inserted at the end of the pipeline and are preceded by three other tools: a sentence splitter, a tokenizer and a POS tagger.

The sentence splitter marks sentence and paragraph boundaries and unwraps sentences split over different lines. An f-score of 99.94% was obtained when testing it on a 12,000 sentence corpus.

The tokenizer segments the text into lexically relevant tokens, using whitespace as the separator; expands contractions; marks spacing around punctuation or symbols; detaches clitic pronouns from the verb; and handles ambiguous strings (contracted vs. non contracted). This tool achieves an f-score of 99.72%.

The POS tagger assigns a single morphosyntactic tag to every token. This tagger is based on Hidden Markov Models, and was developed with the TnT software (Brants, 2000). It scores an accuracy of 96.87%.

2.6 LX-NER

The LX-NER is an online service for the recognition of expressions for named entities in Portuguese. It takes a segment of Portuguese text and identifies, circumscribes and classifies the expressions for named entities it contains. Each named entity receives a standard representation.

This service handles two types of expressions, and their subtypes. (i) Number-based expressions: Numbers — arabic, decimal, non-compliant, roman, cardinal, fraction, magnitude classes; Measures — currency, time, scientific units; Time — date, time periods, time of the day; Addresses — global section, local section, zip code; (ii) Name-base expressions: Persons; Organizations; Locations; Events; Works; Miscellaneous.

The number-based component is built upon handcrafted regular expressions. It was developed and evaluated against a manually constructed test-suite including over 300 examples. It scored 85.19% precision and 85.91% recall. The name-based component is built upon HMMs with the help of TnT (Brants, 2000). It was trained over a manually annotated corpus of approximately 208,000 words, and evaluated against an unseen portion with approximately 52,000 words. It scored 86.53% precision and 84.94% recall.

2.7 CINTIL Concordancer

The CINTIL-Concordancer is an online concordancing service supporting the research usage of the CINTIL Corpus.

The CINTIL Corpus is a linguistically interpreted corpus of Portuguese. It is composed of 1 Million annotated tokens, each one of which verified by human expert annotators. The annotation comprises information on part-of-speech, lemma and inflection of open classes, multi-word expressions pertaining to the class of adverbs and to the closed POS classes, and multi-word proper names (for named entity recognition).

This concordancer permits to search for occurrences of strings in the corpus and returns them together with their window of left and right context. It is possible to search for orthographic forms or through linguistic information encoded in their tags. This service offers several possibilities with respect to the format for displaying the outcome of a given search (e.g. number of occurrences per page, size of the context window, sorting the results in a given page, hiding the tags, etc.)

This service is supported by Poliqarp, a free suite of utilities for large corpora processing (Janus and Przepiórkowski, 2006).

2.8 MWN.PT Browser

The MWN.PT Browser is an online service to browse the MultiWordnet of Portuguese.

The MWN.PT is a lexical semantic network for the Portuguese language, shaped under the ontological model of wordnets, developed by our group. It spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy. These concepts are made of over 21,000 word senses/word forms and 16,000 lemmas from both European and American variants of Portuguese. They are aligned with the translationally equivalent concepts of the English Princeton WordNet and, transitively, of the MultiWordNets of Italian, Spanish, Hebrew, Romanian and Latin.

It includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works, which are covered by the top ontology made of the Portuguese equivalents to all concepts in the 4 top layers of the Princeton wordnet and to the 98 Base Concepts suggested by the Global Wordnet Association, and the 164 Core Base Concepts indicated by the EuroWordNet project.

This browsing service offers an access point to the MultiWordnet, browser[1] tailored to the Portuguese wordnet. It offers also the possibility to navigate the Portuguese wordnet diagrammatically by resorting to Visuwords.[2]

3 Outline

This is an outline of the script to be followed.

Step 1 : Presentation of the LX-Center.
Narrative: The text in Section 2.1 above.
Action: Displaying the page at
`http://lxcenter.di.fc.ul.pt`.

Step 2 : Presentation of LX-Conjugator.
Narrative: The text in Section 2.2 above.
Action: Running an example by selecting ”see an example” option at the page
`http://lxconjugator.di.fc.ul.pt`.

Step 3 : Presentation of LX-Lemmatizer.
Narrative: The text in Section 2.3 above.
Action: Running an example by selecting ”see an example” option at the page
`http://lxlemmatizer.di.fc.ul.pt`;
clicking on one of the inflected forms in the conjugation table generated; clicking on one of the lemmas returned.

Step 4 : Presentation of LX-Inflector.
Narrative: The text in Section 2.4 above.
Action: Running an example by selecting ”see an example” option at the page
`http://lxinflector.di.fc.ul.pt`.

Step 5 : Presentation of LX-Suite.
Narrative: The text in Section 2.5 above.
Action: Running an example by selecting ”see an example” option at the page
`http://lxsuite.di.fc.ul.pt`.

Step 6 : Presentation of LX-NER.
Narrative: The text in Section 2.6 above.
Action: Running an example by copying one of the examples in the page
`http://lxner.di.fc..ul.pt`
and hitting the ”Recognize” button.

Step 7 : Presentation of CINTIL Concordancer.
Narrative: The text in Section 2.7 above.
Action: Running an example by selecting ”see an example” option at the page
`http://cintil.ul.pt`.

Step 8 : Presentation of MWN.PT Browser.
Narrative: The text in Section 2.8 above.
Action: Running an example by selecting ”see an example” option at the page
`http://mwnpt.di.fc.ul.pt/`.

4 Requirements

This demonstration requires a computer (a laptop we will bring along) and an Internet connection.

References

A. Branco, F. Costa, P. Martins, F. Nunes, J. Silva and S. Silveira. 2008. ”LXService: Web Services of Language Technology for Portuguese”. *Proceedings of LREC2008.* ELRA, Paris.

D. Janus and A. Przepiórkowski. 2006. ”POLIQARP 1.0: Some technical aspects of a linguistic search engine for large corpora”. *Proceedings PALC 2005.*

T. Brants. 2000. ”TnT-A Statistical Part-of-speech Tagger”. *Proceedings ANLP2000.*

[1] `http://multiwordnet.itc.it/`

[2] `http://www.visuwords.com/`

A Tool for Deep Semantic Encoding of Narrative Texts

David K. Elson
Columbia University
New York City
delson@cs.columbia.edu

Kathleen R. McKeown
Columbia University
New York City
kathy@cs.columbia.edu

Abstract

We have developed a novel, publicly available annotation tool for the semantic encoding of texts, especially those in the narrative domain. Users can create formal propositions to represent spans of text, as well as temporal relations and other aspects of narrative. A built-in natural-language generation component regenerates text from the formal structures, which eases the annotation process. We have run collection experiments with the tool and shown that non-experts can easily create semantic encodings of short fables. We present this tool as a stand-alone, reusable resource for research in semantics in which formal encoding of text, especially in a narrative form, is required.

1 Introduction

Research in language processing has benefited greatly from the collection of large annotated corpora such as Penn PropBank (Kingsbury and Palmer, 2002) and Penn Treebank (Marcus et al., 1993). Such projects typically involve a formal model (such as a controlled vocabulary of thematic roles) and a corpus of text that has been annotated against the model. One persistent tradeoff in building such resources, however, is that a model with a wider scope is more challenging for annotators. For example, part-of-speech tagging is an easier task than PropBank annotation. We believe that careful user interface design can alleviate difficulties in annotating texts against deep semantic models. In this demonstration, we present a tool we have developed, SCHEHERAZADE, for deep annotation of text.[1]

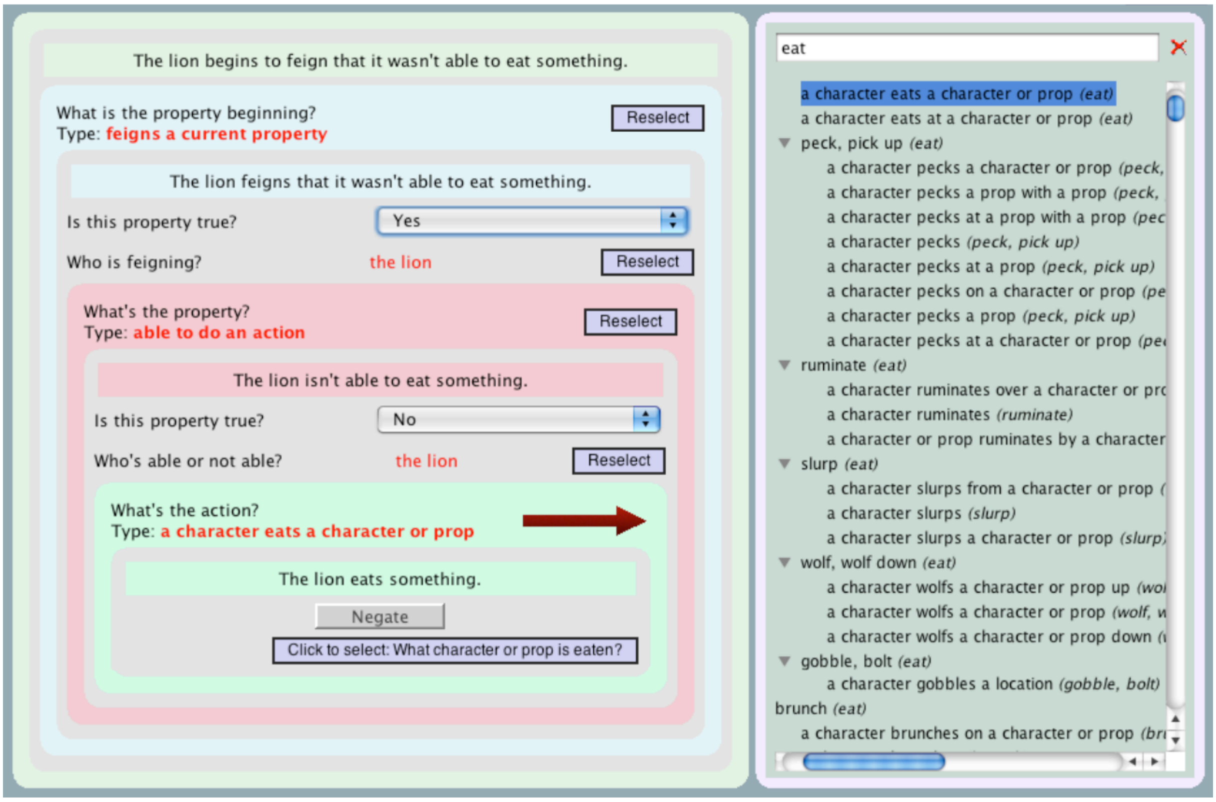
We are using the tool to collect semantic representations of narrative text. This domain occurs frequently, yet is rarely studied in computational linguistics. Narrative occurs with every other discourse type, including dialogue, news, blogs and multi-party interaction. Given the volume of narrative prose on the Web, a system competent at understanding narrative structures would be instrumental in a range of text processing tasks, such as summarization or the generation of biographies for question answering.

In the pursuit of a complete and connected representation of the underlying facts of a story, our annotation process involves the labeling of verb frames, thematic roles, temporal structure, modality, causality and other features. This type of annotation allows for machine learning on the thematic dimension of narrative – that is, the aspects that unite a series of related facts into an engaging and fulfilling experience for a reader. Our methodology is novel in its synthesis of several annotation goals and its focus on content rather than expression. We aim to separate the narrative's *fabula*, the content dimension of the story, from the rhetorical presentation at the textual surface (*sjužet*) (Bal, 1997). To this end, our model incorporates formal elements found in other discourse-level annotation projects such as Penn Discourse Treebank (Prasad et al., 2008) and temporal markup languages such as TimeML (Mani and Pustejovsky, 2004). We call the representation a *story graph*, because these elements are embodied by nodes and connected by arcs that represent relationships such as temporal order and motivation.

More specifically, our annotation process involves the construction of propositions to best approximate each of the events described in the textual story. Every element of the representation is formally defined from controlled vocabularies: the verb frames, with their thematic roles, are adapted from VerbNet (Kipper et al., 2006), the largest verb lexicon available in English. When the verb frames are filled in to construct action

[1] Available at http://www.cs.columbia.edu/~delson.

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 9–12,
Suntec, Singapore, 3 August 2009.



The lion begins to feign that it wasn't able to eat something.

What is the property beginning?
Type: feigns a current property

The lion feigns that it wasn't able to eat something.

Is this property true? Yes

Who is feigning? the lion

What's the property?
Type: able to do an action

The lion isn't able to eat something.

Is this property true? No

Who's able or not able? the lion

What's the action?
Type: a character eats a character or prop

The lion eats something.

eat

a character eats a character or prop *(eat)*
a character eats at a character or prop *(eat)*
▼ peck, pick up *(eat)*
a character pecks a character or prop *(peck,*
a character pecks a prop with a prop *(peck,*
a character pecks at a prop with a prop *(pec*
a character pecks *(peck, pick up)*
a character pecks at a prop *(peck, pick up)*
a character pecks on a character or prop *(pe*
a character pecks a prop *(peck, pick up)*
a character pecks at a character or prop *(pe*
▼ ruminate *(eat)*
a character ruminates over a character or pro
a character ruminates *(ruminate)*
a character or prop ruminates by a character
▼ slurp *(eat)*
a character slurps from a character or prop *(*
a character slurps *(slurp)*
a character slurps a character or prop *(slurp)*
▼ wolf, wolf down *(eat)*
a character wolfs a character or prop up *(wo*
a character wolfs a character or prop *(wolf, w*
a character wolfs a character or prop down *(*
▼ gobble, bolt *(eat)*
a character gobbles a location *(gobble, bolt)*
brunch *(eat)*
a character brunches on a character or prop *(br*

Figure 1: Screenshot from our tool showing the process of creating a formal proposition. On the left, the user is nesting three action propositions together; on the right, the user selects a particular frame from a searchable list. The resulting propositions are regenerated in rectangular boxes.

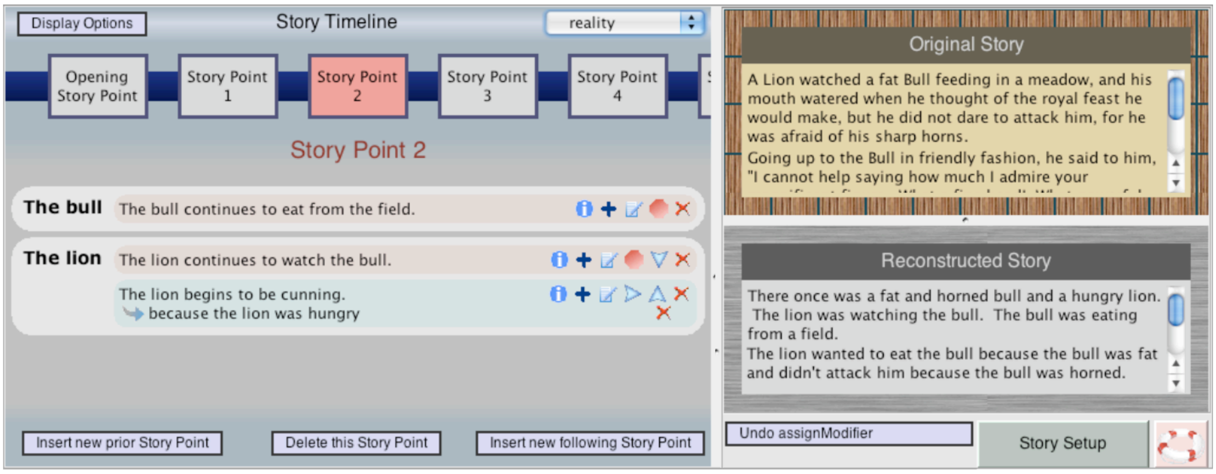
propositions, the arguments are either themselves propositions or noun synsets from WordNet (the largest available noun lexicon (Fellbaum, 1998)). Annotators can also write stative propositions and modifiers (with adjectives and adverbs culled from WordNet), and distinguish between goals, plans, beliefs and other "hypothetical" modalities. The representation supports connectives including causality and motivation between these elements. Finally, and crucially, each proposition is bound to a state (time slice) in the story's main timeline (a linear sequence of states). Additional timelines can represent multi-state beliefs, goals or plans. In the course of authoring actions and statives, annotators create a detailed temporal framework to which they attach their propositions.

2 Description of Tool

The collection process is amenable to community and non-expert annotation by means of a graphical encoding tool. We believe this resource can serve a range of experiments in semantics and human text comprehension.

As seen in Figure 1, the process of creating a proposition with our tool involves selecting an appropriate frame and filling the arguments indicated by the thematic roles of the frame. Annotators are guided through the process by a natural-language generation component that is able to realize textual equivalents of all possible propositions. A search in the interface for "flatter," for example, offers a list of relevant frames such as <*A character*> *flatters* <*a character*>. Upon selecting this frame, an annotator is able to supply arguments by choosing actors from a list of declared characters. "The fox flatters the crow," for one, would be internally represented with the proposition <flatters>([Fox$_1$], [Crow$_1$]) where *flatters*, *Fox* and *Crow* are not snippets of surface text, but rather selected WordNet and VerbNet records. (The subscript indicates that the proposition is invoking a particular [Fox] instance that was previously declared.) In this manner an entire story can be encoded.

Figure 2 shows a screenshot from our interface in which propositions are positioned on a timeline to indicate temporal relationships. On the right side of the screen are the original text (used for reference) and the entire story as regenerated from



The bull continues to eat from the field.

The lion continues to watch the bull.

The lion begins to be cunning.
because the lion was hungry

A Lion watched a fat Bull feeding in a meadow, and his mouth watered when he thought of the royal feast he would make, but he did not dare to attack him, for he was afraid of his sharp horns.

Going up to the Bull in friendly fashion, he said to him, "I cannot help saying how much I admire your

There once was a fat and horned bull and a hungry lion. The lion was watching the bull. The bull was eating from a field.

The lion wanted to eat the bull because the bull was fat and didn't attack him because the bull was horned.

Figure 2: The main screen of our tool features a graphical timeline, as well as boxes for the reference text and the story as regenerated by the system from the formal model.

the current state of the formal model. It is also possible from this screen to invoke modalities such as goals, plans and beliefs, and to indicate links between propositions. Annotators are instructed to construct propositions until the resulting textual story, as realized by the generation component, is as close to their own understanding of the story as permitted by the formal representation.

The tool includes annotation guidelines for constructing the best propositions to approximate the content of the story. Depending on the intended use of the data, annotators may be instructed to model just the stated content in the text, or include the implied content as well. (For example, causal links between events are often not articulated in a text.) The resulting story graph is a unified representation of the entire *fabula*, without a story's beginning or end. In addition, the tool allows annotators to select spans of text and link them to the corresponding proposition(s). By indicating which propositions were stated in the original text, and in what order, the content and presentation dimensions of a story are cross-indexed.

3 Evaluation

We have conducted several formative evaluations and data collection experiments with this interface. In one, four annotators each modeled four of the fables attributed to Aesop. In another, two annotators each modeled twenty fables. We chose to model stories from the Aesop corpus due to several key advantages: the stories are mostly built from simple declaratives, which are within the expressive range of our semantic model, yet are rich in thematic targets for automatic learning (such as dilemmas where characters must choose from between competing values).

In the latter collection, both annotators were undergraduates in our engineering school and native English speakers, with little background in linguistics. For this experiment, we instructed them to only model stated content (as opposed to including inferences), and skip the linking to spans of source text. On average, they required 35-45 minutes to encode a fable, though this decreased with practice. The 40 encodings include 574 propositions, excluding those in hypothetical modalities. The fables average 130 words in length (so the annotators created, on average, one proposition for every nine words).

Both annotators became comfortable with the tool after a period of training; in surveys that they completed after each task, they gave Likert-scale usability scores of 4.25 and 4.30 (averaged over all 20 tasks, with a score of 5 representing "easiest to use"). The most frequently cited deficiencies in the model were abstract concepts such as *fair* (in the sense of a community event), which we plan to support in a future release.

4 Results and Future Work

The end result from a collection experiment is a collection of story graphs which are suitable for machine learning. An example story graph, based on the state of the tool seen in Figure 2, is shown in Figure 3. Nodes in the graph represent states, declared objects and propositions (actions and statives). Each of the predicates (e.g., <lion>,

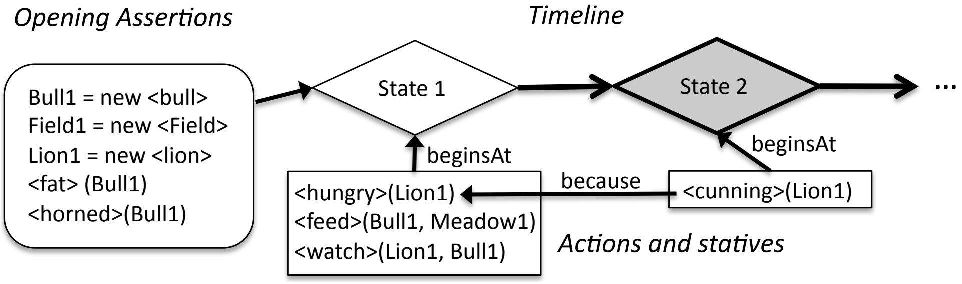


Figure 3: A portion of a story graph representation as created by SCHEHERAZADE.

<watch>, <cunning>) are linked to their corresponding VerbNet and WordNet records.

We are currently experimenting with approaches for data-driven analysis of narrative content along the “thematic” dimension as described above. In particular, we are interested in the automatic discovery of deep similarities between stories (such as analogous structures and prototypical characters). We are also interested in investigating the selection and ordering of content in the story’s telling (that is, which elements are stated and which remain implied), especially as they pertain to the reader’s affectual responses. We plan to make the annotated corpus publicly available in addition to the tool.

Overall, while more work remains in expanding the model as well as the graphical interface, we believe we are providing to the community a valuable new tool for eliciting semantic encodings of narrative texts for machine learning purposes.

5 Script Outline

Our demonstration involves a walk-through of the SCHEHERAZADE tool. It includes:

1. An outline of the goals of the project and the innovative aspects of our formal representation compared to other representations currently in the field.

2. A tour of the timeline screen (equivalent to Figure 2) as configured for a particular Aesop fable.

3. The procedure for reading a text for important named entities, and formally declaring these named entities for the story graph.

4. The process for constructing propositions in order to encode actions and statives in the text, as seen in Figure 1.

5. Other features of the software package, such as the setting of causal links and the ability to undo/redo.

6. A review of the results of our formative evaluations and data collection experiments, including surveys of user satisfaction.

References

Mieke Bal. 1997. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press, Toronto, second edition.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02)*, Canary Islands, Spain.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extensive classifications of english verbs. In *Proceedings of the 12th EURALEX International Congress*, Turin, Italy.

Inderjeet Mani and James Pustejovsky. 2004. Temporal discourse models for narrative structure. In *Proceedings of the ACL Workshop on Discourse Annotation*, Barcelona, Spain.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

ProLiV - a Tool for Teaching by Viewing Computational Linguistics

Monica Gavrila
Hamburg University, NATS
Vogt-Kölln Str 30, 20251, Germany
gavrila@informatik.uni-hamburg.de

Cristina Vertan
Hamburg University, NATS
Vogt-Kölln Str 30, 20251, Germany
vertan@informatik.uni-hamburg.de

Abstract

ProLiV - Animated Process-modeler of Complex (Computational) Linguistic Methods and Theories - is a fully modular, flexible, XML-based stand-alone Java application, used for computer-assisted learning in Natural Language Processing (NLP) or Computational Linguistics (CL). Having a flexible and extendible architecture, the system presents the students, by means of text, of visual elements (such as pictures and animations) and of interactive parameter set-up, the following topics: Latent Semantics Analysis (LSA), (computational) lexicons, question modeling, Hidden-Markov-Models (HMM), and Topic-Focus. These topics are addressed to first-year students in computer science and/or linguistics.

1 Introduction

The role of multimedia in teaching Natural Language Processing (NLP) is demonstrated by constant development of software packages such as GATE (http://gate.ac.uk) and NLTK (http://nltk.sourceforge.net/index.html). Detailed information about visual tools for NLP, in particular about GATE, is to be found in (Gaizauskas et al, 2001).

ProLiV is a Java application framework, developed in a three-year project (2005-2008) at the University of Hamburg. It helps first-year students to understand and learn, in an easier manner, either complex linguistic theories used in NLP (e.g. question modeling) or statistical approaches for computational linguistics (e.g. LSA, HMM).

The learning process is supported by modules integrating text, visual and interactive elements. In its first released version, ProLiV contains the following modules:

- the Latent Semantic Analysis (LSA) module and the computational lexicons module - for linguists,
- the question modeling module - for computer scientists,
- the Hidden-Markov-Models (HMM) module and Topic-Focus module - for both computer scientists and linguists.

2 The Learning Path

For each module, the learning path is guided by lessons, a terminology dictionary and interactive activities. Exercises and small tests can also be integrated.

The lessons include text, pictures and animations. Hyperlinks between lessons ensure a concept-oriented navigation through the learning content. Additionally key terms within the content are linked with dictionary entries.

Three central issues guided the development of the ProLiV software:

1. choosing the most adequate means (text / picture / animation) to represent lessons content,
2. designing the layout (quantity and size of text, colors) in order to increase the learning success,
3. in case of the animations, defining its components and parameters (speed, animation steps, and graphical elements) to maximize their impact on users.

Regarding the second issue above-mentioned, the layout of the modules follows part of the guidelines found in (Orr et al., 1994) and (Thibodeau, 1997).

Considering the current multimedia development, the trend is using animations to improve the learning process. Animations are assumed to be

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 13–16,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP

a promising educational tool, although their efficiency is not fully proved. Researchers, such as (Morrison, 2000), showed that animations can convey more information and be helpful when showing details in intermediate steps of a process, but when building an animation it is very important to consider the background of the student (e.g. linguistics, natural sciences) and his/her psychological functioning. The educational effectiveness of the animations depends on how they interact with the learner. Depending on the student's background, in order to have a helpful material, one has to carefully decide what information the animation contains. As our experiment showed (see Section 2.1), depending on the student and his/her background, an animation can improve the learning process, or bring nothing to it. We found no cases when the animation slowed down the learning process.

The system was experimentally used in seminars at the University of Hamburg. Part of the lessons content was adapted following the user's feedback.

2.1 Animations in ProLiV

Animations are not integrated in all modules of the ProLiV system, but only in the LSA, computational lexicons and question modeling modules.

In order to decide how to organize the information in an animation, we evaluated the animations for the matrix multiplication in the LSA module by asking 11 high-school pupils (between 16 and 19 years old) to choose between the several representations.

We showed the pupils three animations that describe the multiplication of matrices, a static picture and the text representation of the definition. The animations differ in the way the process is presented (abstract vs. concrete) and in user interaction authorization.

The pupils were asked to evaluate all the representations. The question they had to answer was: "*Which of the following representations helps more, when learning about matrix multiplication?*". The scale given was from 1 = very helpful to 5 = not helpful at all.

Analyzing the results, we could not conclude that one representation is a "*real winner*'. The best representation was considered the most flexible animation, that allows the student go backwards and forwards whenever the user needs it, the learning process being adapted to the user's rhythm. All the evaluation results can be seen in Table 1. In order to better see the influence of these representations in the learning process, statistical tests should be run.

Representation	Average Result
Definition (formula)	3.5
Picture	2.91
Animation 1	3.64
Animation 2	**2.09**
Animation 3	2.45

Table 1: Evaluation of the animations in the matrix multiplication (*Animations 1 and 3 have no user interaction; Animations 1 and 2 are more abstract*)

3 System Architecture

In Figure 1 we present the ProLiV System architecture, consisting of:

- a file repository (lessons, dictionary, tests, and exercises),
- a tool repository,
- an aggregating module combining elements from file and tool repository (Main Unit),
- the graphical user interface (G.U.I.)

For each topic a stand-alone module is connected with the G.U.I module via the Main Unit. Modules related to new topics can be inserted any time with no particular changes of the system.

The ProLiV architecture follows the guideline considerations found in (Galitz, 1997).

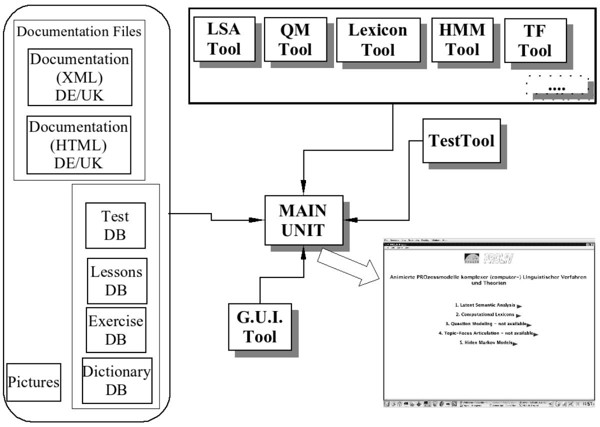

Documentation Files
Documentation (XML) DE/UK
Documentation (HTML) DE/UK
Test DB
Lessons DB
Exercise DB
Dictionary DB
Pictures
LSA Tool
QM Tool
Lexicon Tool
HMM Tool
TF Tool
....
TestTool
MAIN UNIT
G.U.I. Tool
PROLIV
Animierte PROzessmodelle komplexer (computer-) Linguistischer Verfahren und Theorien
1. Latent Semantic Analysis
2. Computational Lexicons
3. Question Modeling - not available
4. Topic-Focus Articulation - not available
5. Hiden Markov Models

Figure 1: The ProLiV Architecture

The flexibility of the system is also given by the fact that the G.U.I.[1] is generated according to an XML[2] description, developed within the project (see DTD Description).

The XML description contains the information in the lessons (definitions, theory, examples, etc.) and the G.U.I. specifications (colors, fonts, links, arrangement in the interface, etc.). Having an XML file as input, the system generates automatically the G.U.I. presented to the student. The information shown to the user can be extended or modified with almost no implementation effort. New lessons or modules can be integrated, by extending or adding XML files. Due to the same fact, also the content adaptation of the system to other languages[3] is very easy.

```
The DTD Description:

<?xml version=’’1.0’’?>
<DOCTYPE LESSONS[
<!ELEMENT LESSONS (LESSON+)>
<!ELEMENT LESSON (TITLE+, (TEXT|FORMULA|
INDEXI|INDEX|BOLD|
ITALIC|TERM|LINK|DEF|
EXM|OBS|T|OTHER)+>
<!ELEMENT TITLE (#PCDATA)>
<!ELEMENT TEXT (#PCDATA)>
<!ELEMENT FORMULA (#PCDATA)>
<!ELEMENT INDEX (#PCDATA)>
<!ELEMENT INDEXI (#PCDATA)>
<!ELEMENT BOLD (#PCDATA)>
<!ELEMENT ITALIC (#PCDATA)>
<!ELEMENT TERM (#PCDATA)>
........................
<!ELEMENT T (#PCDATA)>
<!ELEMENT OTHER (#PCDATA)>

<!ATTLIST LESSON NO CDATA #REQUIRED>
<!ATTLIST DEF NO CDATA #REQUIRED>
<!ATTLIST EXM NO CDATA #REQUIRED>
<!ATTLIST OBS NO CDATA #REQUIRED>
<!ATTLIST QUIZZ NO CDATA #REQUIRED>
<!ATTLIST EX NO CDATA #REQUIRED>
<!ATTLIST T NO CDATA #REQUIRED>
<!ATTLIST OTHER STYLE CDATA #REQUIRED>
```

The G.U.I. follows the same design rules in all modules and the layout and format decisions are consistent. A color and a font style are associated to only one kind of information (e.g. color red associated to definitions, etc.).

3.1 Integrated external software packages

The learning process is also sustained by interactive elements, such as the possibility of changing parameters for the LSA algorithm and visualizing the results, or as the integrated programs for the computational lexicons tool: ManageLex (`http://nats-www.informatik.uni-hamburg.de/view/Main/ManageLex`) and G.E.R.L. (`http://nats-www.informatik.uni-hamburg.de/view/Main/GerLexicon`). This way the students have the possibility, not only to read the theory, but also to see the impact of their modifications in an algorithm that is described in the lessons.

Due to its architecture, other such external programs can be easily integrated within ProLiV.

4 LSA Module in ProLiV

In order to have a better overview of what a module contains and how it is organized, this section presents some aspects of the LSA module.

The LSA module makes an introduction to the topic. It gives an overview of the LSA algorithm, principles, application areas, and of the main mathematical notions used in the algorithm. Initially thought for being used mostly by students from linguistics (or linguists) - due to the mathematical algorithms -, the tool can be exploited by anybody who wants to have an introductory course on LSA.

The content is organized in four Units:

1. LSA: General Knowledge - It gives the LSA definition, a short overview of the history, its semantics, and how LSA can be used in the study of cognitive processes.

2. Mathematical Fundamentals - It describes the LSA algorithm

3. LSA Applications - It presents the application areas for the LSA, LSA limitations and critics. Also a comparison with other similar algorithms is made.

4. Compendium of Mathematics - It gives the user the mathematical background: definitions, theorems, etc.

The course has also an introduction, a motivation, conclusion and references.

[1] The G.U.I. is automatically generated not only for the lessons, but also for the term dictionary associated to each module.

[2] XML = Extensible Markup Language. More details to be found on `http://en.wikipedia.org/wiki/XML`

[3] For the moment ProLiV contains lessons in German and English

The LSA module is offering not only a textual representation of the information, but also several visualization methods (as images and animations[4]). Beside the lessons, there are implemented a term dictionary and an environment for testing LSA parameters.

4.1 The LSA Test Environment

Probably the most interesting part of the LSA module is the test environment. After learning about LSA, in this environment the user has the possibility to actually see how LSA is working, and what results can be obtained when comparing the meaning of two words. The user can set several parameters of the algorithm - e.g. the analysis mode (simple/frequency based vs. advanced/entropy based), the minimum word occurrences, the analysis dimension, the similarity measure (Cosine, Euclidean, Pearson, Dot-Product), etc. - and decide which words are not considered in the analysis. The analyzed text, the initial cooccurrence matrix and the one obtained after applying the Singular Value Decomposition (SVD) algorithm are shown in the G.U.I. The similarity measure, when comparing two words, is calculated in both unreduced and reduced cases.

5 Conclusions

The paper presents a course-ware software, ProLiV. It is a collection of (interactive) multimedia tools used mainly for the consolidation of first-years courses in computational linguistics and literary computing. Its goal is to help the humanist scientists to make use of complex formal methods, and the computer specialists to understand humanist facts and interpretations.

The main feature of the system, in the context of the conference, is not the content of the lessons, but the system's extendible and adaptable architecture. Another important aspect is the way in which the information is presented to the student.

The system runs on any platform supporting Java 1.5 or newer. It was developed on Linux and tested on Windows and Mac OS X.

Being Java-based and having as input Unicode files (XML encoded information), the system can be embedded in the future in a Web environment.

More about ProLiV can be found in (Gavrila et al, 2006) or in (Gavrila et al, TBA) and on the ProLiV homepage: `http://nats-www.informatik.uni-hamburg.de/view/PROLIV/WebHome`.

Acknowledgments

We would like to thank all people that helped in the development of our software: Project Coordinator Prof. Dr. Walther v. Hahn (Computer Science Department, Natural Language Systems Group), Prof. Dr. Angelika Redder (Department of Language, Linguistics and Media Studies, Institute for German Studies I), Dr. Shinichi Kameyama (Department of Language, Linguistics and Media Studies, Institute for German Studies I), Christina von Bremen (Computer Science Department, Natural Language Systems Group), Olga Szczepanska (Computer Science Department, Natural Language Systems Group), Irina Aleksenko (Computer Science Department, Natural Language Systems Group), Svetla Boytcheva (Academy of Sciences Sofia).

References

Wilbert O. Galitz. 1997 *The Essential Guide to User Interface Design: an Introduction to GUI Design principles and Techniques*, Wiley Computer Publishing, New York.

Robert J. Gaizauskas, Peter J. Rodgers, and Kevin Humphreys. 2001 *Visual Tools for Natural Language Processing*, *Journal of Visual Languages and Computing*, Vol. 12, Number 4, p. 375-411, Academic Press

Monica Gavrila, Cristina Vertan. 2006 *Visualization of Complex Linguistic Theories*, in the Proceedings of the ICDML 2006 Conference, p. 158-163, Bangkok, Thailand, March 13-14

Monica Gavrila, Cristina Vertan, and Walther von Hahn. To be published during 2009 *ProLiV - Learning Terminology with animated Models for Visualizing Complex Linguistics Theories*, in the Proceedings of the LSP 2007 Conference, Hamburg, Germany, August,

Julie Bauer Morrison, Barbara Twersky, and Mireille Betrancourt. 2000 *Animation: Does It Facilitate Learning?*, in the Proc. of the Workshop on Smart Graphics, AAAI Press, Menlo Park, CA.

Kay L .Orr, Katharine C. Golas, and Katy Yao. 1994 *Storyboard Development for Interactive Multimedia Training*, *Journal of Interactive Instruction Development*, Volume 6, Number 3, p. 18-31

Pete Thibodeau. 1997 *Design Standards for Visual Elements and Interactivity for Courseware*, *T.H.E. Journal*, Volume 24, Number 7, p. 84-86

[4] The animations integrated are for the LSA algorithm tested on an example and for matrix multiplication

A Web-Based Interactive Computer Aided Translation Tool

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

Abstract

We developed **caitra**, a novel tool that aids human translators by (a) making suggestions for sentence completion in an interactive machine translation setting, (b) providing alternative word and phrase translations, and (c) allowing them to post-edit machine translation output. The tool uses the Moses decoder, is implemented in Ruby on Rails and C++ and delivered over the web.

1 Introduction

Today's machine translation systems are mostly used for inbound translation (also called assimilation), where the reader accepts lower quality translation for instant access to foreign language text. The standards are much higher for outbound translation (also called dissemination), where the reader is typically an unsuspecting customer or citizen who is seeking information about products or services, and human translators are required for high-quality publication-ready translation.

While machine translation has made tremendous progress over the last years, this progress has made little inroads into tools for human translators. Although it has become common practice in the industry to provide human translators with machine translation output that they have to post-edit, typically no deeper integration of machine translation and human translation is found in translation agencies.

An interesting approach was pioneered by the TransType project (Langlais et al., 2000). The machine translation system makes sentence completion predictions in an interactive machine translation setting. The users may accept them or override them by typing in their own translations, which triggers new suggestions by the tool (Barrachina et al., 2009).

But also other information that is generated during the machine translation process may be useful for the human translator, such as alternative translations for the input words and phrases.

We are at the beginning of a research program to explore the benefits of these different types of aid to human translators, analyze user interaction behavior, and develop novel types of assistance. To have a testbed for this research, we developed an online, web-based tool for translators.

2 Overview

Caitra is implemented in Ruby on Rails (Thomas and Hansson, 2008) as a web-based client-server architecture, using Ajax-style Web 2.0 technologies (Raymond, 2007) connected to a MySQL database-driven back-end. The machine translation back-end is powered by the open source Moses decoder (Koehn et al., 2007). The interactive machine translation prediction code is implemented in C++ for speed. The tool is delivered over the web to allow for easier user studies with remote users, but also to expose the tool to a wider community to gather additional feedback. You can find caitra online at `http://www.caitra.org/`

Caitra allows the uploading of documents using a simple text box. This text is then processed by a back-end job to pre-compute all the necessary data (machine translation output, translation options, search graphs). This process takes a few minutes.

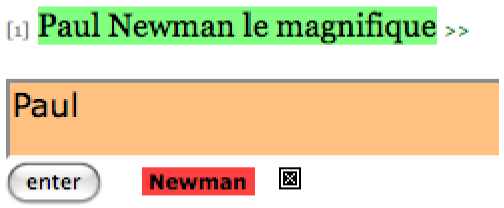
Finally, the user is presented with an interface that includes all the different types of assistance. Each may be turned off, if the user finds it distracting. The user translates one sentence at a time, while the context (both input and user translation, including the proceeding and following paragraph) is displayed for reference.

In the next three sections, we will describe each type of assistance in detail.

3 Interactive Machine Translation

The idea of interactive machine translation has been greatly advanced by work carried out in the TransType project (Langlais et al., 2000), with the focus on a sentence-completion paradigm. While the human translator is still in charge of creating

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 17–20,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP


[1] Paul Newman le magnifique >>

Paul

Figure 1: **Interactive Machine Translation.** Caitra uses the search graph of the machine translation decoder to suggest words and phrases to continue the translation.

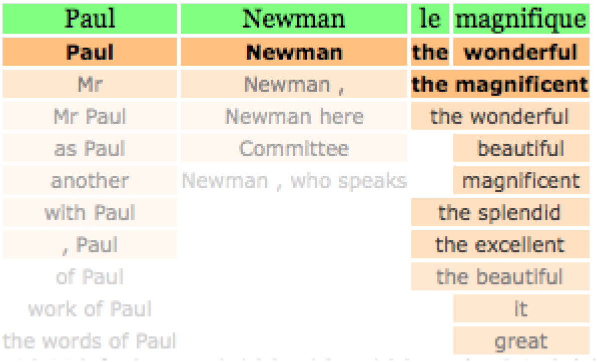
the translation word by word, she is aided by a machine translation system that interactively makes suggestions for completing the sentence, and updates these suggestions based on user input. The scenario is very similar to the auto-completion function for words, search terms, email addresses, etc. in modern office applications.

See Figure 1 for a screenshot of the incarnation of this method in our translation tool. The user is given an input sentence and a standard web text box to type in her translation. In addition, caitra makes suggestions about the next word (or phrase) to be added to the translation. The user may accept this (by pressing the TAB key), or type in her own translation. The tool updates the prediction based on the user input.

The predictions are based on a statistical machine translation system. Given the input and the partial translation of the user (called the prefix), the machine translation system computes the optimal translation of the input sentence, constrained by matching the user input. This translation is provided to the user in form of short phrases (mirroring the underlying phrase-based statistical translation model).

In contrast to traditional work on interactive machine translation, the displayed suggestions consist of only very few words to not overload the reading capacity of the user. We have not yet carried out studies to explore the optimal length of suggestions, or even when not to provide suggestions at all, in cases when they will be most likely useless and distractive.

We store the search graph produced by the machine translation decoder in a database. During the user interaction, we quickly match user input against the graph using a string edit distance measure. The prediction is the optimal completion path that matches the user input with (a) minimal string edit distance and (b) highest sentence translation probability. This computation takes place at the server and is implemented in C++.



Paul	Newman	le	magnifique
Paul	Newman	the	wonderful
Mr	Newman ,	the magnificent	
Mr Paul	Newman here	the wonderful	
as Paul	Committee		beautiful
another	Newman , who speaks		magnificent
with Paul		the splendid	
, Paul		the excellent	
of Paul		the beautiful	
work of Paul			it
the words of Paul			great

Figure 2: **Translation Options.** The most likely word and phrase translation are displayed alongside the input words, ranked and color-coded by their probability.

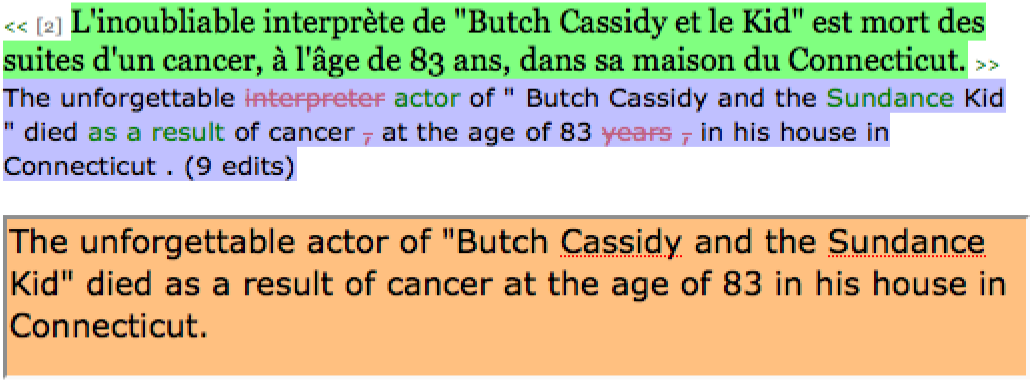
While caitra only displays one phrase prediction at a time, the entire completion path is transmitted to the client. Acceptance of a system suggestion will instantly lead to another suggestion, while typed-in user translations require the computation of a new sentence completion path. This typically takes less than a second.

Preliminary studies suggest that users accept up to 50-80% of system predictions, but obviously this number depends highly on language pair and difficulty of the text.

4 Options from the Translation Table

Phrase-based statistical machine translation methods acquire their translation knowledge in form of large phrase translation tables automatically from large amounts of translated texts (Koehn et al., 2003). For each input word or input word sequence, this translation table is consulted for the most likely translation options. A heuristic beam search algorithm explores these options and their ordering to find the most likely sentence translation (which takes into account various scoring functions, such as the use of an n-gram language model).

These translation options may also be of interest to the user, so we display them in our translation tool caitra. See Figure 2 for an example. For instance, the tool suggests for the translation of the French *magnifique* the English options *wonderful, beautiful, magnificent,* and *great*, among others. The user may click on any of these phrases and



<< [2] L'inoubliable interprète de "Butch Cassidy et le Kid" est mort des suites d'un cancer, à l'âge de 83 ans, dans sa maison du Connecticut. >>

The unforgettable ~~interpreter~~ actor of " Butch Cassidy and the Sundance Kid " died as a result of cancer ~~,~~ at the age of 83 ~~years~~ ~~,~~ in his house in Connecticut . (9 edits)

The unforgettable actor of "Butch Cassidy and the Sundance Kid" died as a result of cancer at the age of 83 in his house in Connecticut.

Figure 3: **Post-Editing Machine Translation.** Starting with the sentence translation of the machine translation system, the user post-edits and the tool indicates changes.

they are added into the text box. The user may also just glance at these suggestions and then type in the translations herself.

The options are color-coded and ranked based on their score. Note that since these options are extracted from a translated corpus using various automatic methods, often inappropriate translations are included, such as the translation of *Newman* into *Committee*.

For each translation option a score is computed to assess its utility. This score is the (i) future cost estimates of the phrases (ii) plus the outside cost estimates for the remaining sentence (iii) minus the future cost estimate for the full sentence. This number allows the ranking of words vs. phrases of different length. The ranking of the phrases never places a lower scoring option above a higher scoring option. The absolute score is used to color code the options. Up to ten table rows are filled with options.

Since the user may click on the options, or may simply type in translations inspired by the options, it is not straight-forward to evaluate their usefulness. We plan to assess this by measuring translation speed and quality. Experience so far has shown that the options help novice users with unknown words and advanced users with suggestions that are not part of their active vocabulary. It may be possible that these options even allow users that do not know the source language to create a translation, as in work done by Albrecht et al. (2009).

5 Post-Editing Machine Translation

The addition of full sentence translation of the machine translation system is trivial compared to the other types of assistance. When a user starts a new sentence using this aid, the text box already contains the machine translation output and the user only makes changes to correct errors.

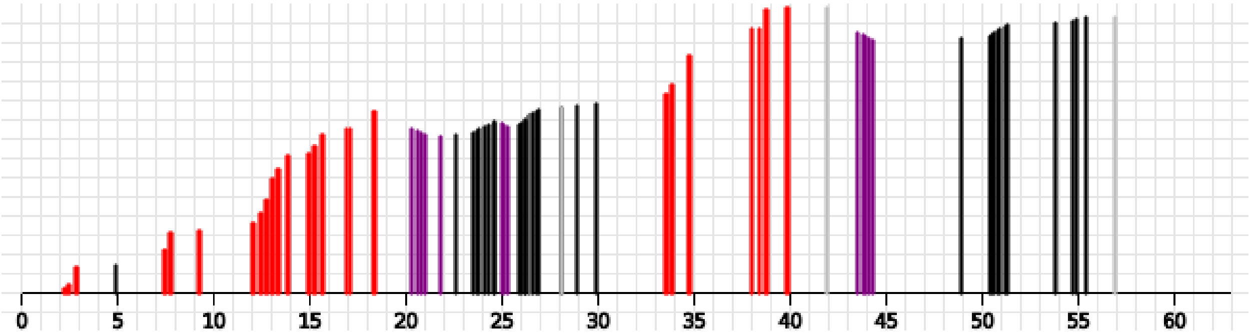
See Figure 3 for an example. Caitra also compares the user's translation in form of string edit distance against the original machine translation. This is illustrated above the text box, to possibly alert the user to mistakenly dropped or added content.

6 Key Stroke Logging

Caitra tracks every key stroke and mouse click of the user, which then allows for a detailed analysis of the user's interaction with the tool. See Figure 4 for a graphical representation of the user activity during the translation of a sentence. The graph plots sentence length (in characters) against the progression of time. Bars indicate the sentence length at each point in time when a user action takes place (acceptance of predictions are red, DEL key strokes purple, key strokes for cursor movement grey, and key strokes that add characters are black.)

In the example sentence, the user first slowly accepted the interactive machine translation predictions (second 0-12), then more rapidly (second 12-20), followed by a period of deletions and typing that did not make the translation longer (second 20-30). After a short pause, predictions were accepted again (second 33-40), followed by deletions and typing (second 40-57).

We are currently carrying out user studies to not only compare the productivity improvements gained by the different types of help offered to the user, but also to identify, categorize and analyze the types of activities (such as long pauses,



Input: ”Un échange de coups de feu s’est produit, et la moitié des ravisseurs ont été tués, les autres s’enfuyant”, a dit ce responsable qui a requis l’anonymat.

MT: ”A exchange of fire occurred, and half of the kidnappers were killed, the other is enfuyant,” said this official who has requested anonymity.

User: ”An exchange of fire occurred, and half of the kidnappers were killed, the others running away”, said the source who has requested anonymity.

Figure 4: **User Activity.** The graph plots the time spent on translation (in seconds, x-axis) against the length of the sentence (y-axis) with color-coded activities (bars). For instance, at the interval second 2–3, three interactive machine translations predictions were accepted.

slow typing, fast typing, clicks on options, acceptance of predictions) to gain insight into the type of problems in (computer aided) human translation and the time spent to solve these problems.

7 Conclusions

We described the new computer aided translation tool caitra that allows us to compare industry-standard post-editing, the interactive sentence completion paradigm, and other help for translators. The tool is available online at the URL `http://www.caitra.org/`.

We will report on user studies in future papers.

8 Acknowledgments

This work was supported by the EuroMatrix-Plus project funded by the Europea Commission (7th Framework Programme). Thanks to Josh Schroeder for help with Ruby on Rails.

References

Albrecht, J., Hwa, R., and Marai, G. E. (2009). Correcting automatic translations through collaborations between mt and monolingual target-language users. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Langlais, P., Foster, G., and Lapalme, G. (2000). Transtype: a computer-aided translation typing system. In *Proceedings of the ANLP-NAACL 2000 Workshop on Embedded Machine Translation Systems*.

Raymond, S. (2007). *Ajax on Rails*. O’Reilly.

Thomas, D. and Hansson, D. H. (2008). *Agile Web Development with Rails: Second Edition, 2nd Edition*. The Pragmatic Programmers, LLC.

MARS: Multilingual Access and Retrieval System with Enhanced Query Translation and Document Retrieval

Lianhau Lee, Aiti Aw, Thuy Vu, Sharifah Aljunied Mahani, Min Zhang, Haizhou Li
Institute for Infocomm Research
1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
{lhlee, aaiti, tvu, smaljunied, mzhang, hli}
@i2r.a-star.edu.sg

Abstract

In this paper, we introduce a multilingual access and retrieval system with enhanced query translation and multilingual document retrieval, by mining bilingual terminologies and aligned document directly from the set of comparable corpora which are to be searched upon by users. By extracting bilingual terminologies and aligning bilingual documents with similar content prior to the search process provide more accurate translated terms for the in-domain data and support multilingual retrieval even without the use of translation tool during retrieval time. This system includes a user-friendly graphical user interface designed to provide navigation and retrieval of information in browse mode and search mode respectively.

1 Introduction

Query translation is an important step in the cross-language information retrieval (CLIR). Currently, most of the CLIR system relies on various kinds of dictionaries, for example WordNets (Luca and Nurnberger, 2006; Ranieri et al., 2004), in query translation. Although dictionaries can provide effective translation on common words or even phrases, they are always limited in the coverage. Hence, there is a need to expand the existing collections of bilingual terminologies through various means.

Recently, there has been more and more research work focus on bilingual terminology extraction from comparable corpora. Some promising results have been reported making use of statistics, linguistics (Sadat et al., 2003), transliteration (Udupa et al., 2008), date information (Tao and Zhai, 2005) and document alignment approach (Talvensaari et al., 2007).

In this paper, we introduce our Multilingual Access and Retrieval System – MARS which addresses the query translation issue by using in-domain bilingual terminologies extracted directly from the comparable corpora which are to be accessed by users. And at the same time, bilingual documents are paired up prior to the search process based on their content similarities to overcome the limitation of traditional keyword matching based on the translated terms. These would provide better retrieval experiences as not only more accurate in-domain translated term will be used to retrieve the documents but also provide a new perspective of multilingual information retrieval to process the time-consuming multilingual document matching at the backend.

The following sections of this paper will describe the system architecture and the proposed functionalities of the MARS system.

2 MARS System

The MARS system is designed to enhance query translation and document retrieval through mining the underlying multilingual structures of comparable corpora via a pivot language. There are three reasons for using a pivot language. Firstly, it is appropriate to use a universal language among potential users of different native languages. Secondly, it reduces the backend data processing cost by just considering the pair-wise relationship between the pivot language and any other languages. Lastly, the dictionary resources between the pivot language and all the other languages are more likely to be available than otherwise.

There are two main parts in this system, namely data processing and user interface. The data processing is an offline process to mine the underlying multilingual structure of the compa-

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 21–24,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP

rable corpora to support retrieval. The structure of the comparable corpora is presented visually in the user interface under browse mode and search mode to facilitate navigation and retrieval of information respectively.

3 Data Processing

For demo purpose, three different language newspapers from the year 1995 to 2006 published by Singapore Press Holding (SPH), namely Strait Times[1] (English), ZaoBao[2] (Chinese) and Berita Harian[3] (Malay), are used as comparable corpora. In these particular corpora, English is chosen as the pivot language and noun terms are chosen as the basic semantic unit as they represent a huge amount of significant information. Our strategy is to organize and manipulate the corpora in three levels of abstraction – clusters, documents and terms. And our key task over here is to find the underlying associations of documents or terminologies in each level across different languages.

First, monolingual documents are grouped into clusters by k-means algorithm using simple word vectors. Then, monolingual noun terms are extracted from each cluster using linguistic patterns and filtered by occurrence statistics globally (within cluster) and locally (within document), so that they are good representatives for cluster as a whole as well as individual documents (Vu et al., 2008). The extracted terms are then used in document clustering in a new cycle and the whole process is repeated until the result converges.

Next, cluster alignment is carried out between the pivot language (English) and the other languages (Chinese, Malay). Clusters can be conceptualized as the collection of documents with the same themes (e.g. finance, politics or sports) and their alignments as the correspondents in the other languages. Since there may be overlaps among themes, e.g. finance and economy, each cluster is allowed to align to more than one cluster with varying degree of alignment score.

After that, document alignment is carried out between aligned cluster pairs (Vu et al., 2009). Note that the corpora are comparable, thus the aligned document pairs are inherently comparable, i.e. they are similar in contents but not identical as translation pairs. Also as important to note that, document alignment harvested over here is independent of user query. In other words, document alignment is not simply determined by mere occurrence of certain keyword and its absence does not hinder documents to be aligned. Hence mining of document alignment beforehand improves document retrieval afterward.

Finally, term alignment is likewise generated between aligned document pairs. The aligned terms are expected to be in-domain translation pairs since they are both derived from documents of similar contents, and thus they have similar contexts. By making use of the results provided by each other, document alignment and term alignment can be improved over iterations.

All the mentioned processes are done offline and the results are stored in a relational database which will handle online queries generated in the user interface later on.

4 User Interface

As mentioned, there are two modes provided in the user interface to facilitate navigation and retrieval of information, namely browse mode and search mode. Both modes can be switched simply by clicking on the respective tabs in the user interface. In the following, the functionalities of the browse mode and the search mode will be explained in details.

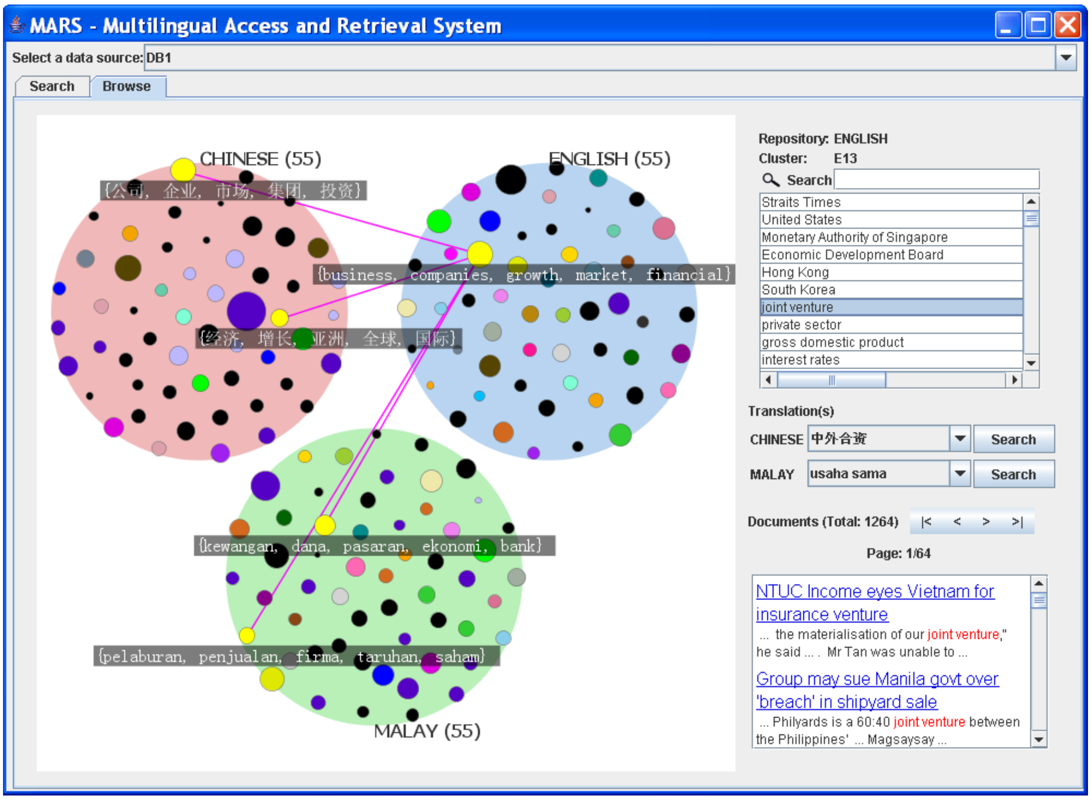
4.1 Browse Mode

Browse mode provides a means to navigate through the complex structures underneath an overwhelming data with an easily-understood, user-friendly graphical interface. In the figure 1, the graph in the browse mode gives an overall picture of the distribution of documents in various clusters and among the different language collections. The outer circles represent the language repositories and the inner circles represent the clusters. The sizes of the clusters are depending on the number of contained documents and the color represents the dominant theme. The labels of the highlighted clusters, characterized by a set of five distinguished words, are shown in the tooltips next to them. By clicking on a cluster, the links depicting the cluster alignments will show up. The links to the clusters in the other languages are all propagated through the pivot language.

[1] http://www.straitstimes.com/ an English news agency in Singapore. Source © Singapore Press Holdings Ltd.

[2] http://www.zaobao.com/ a Chinese news agency in Singapore. Source © Singapore Press Holdings Ltd.

[3] http://cyberita.asia1.com.sg/ a Malay news agency in Singapore. Source © Singapore Press Holdings Ltd.


MARS - Multilingual Access and Retrieval System
Search
Browse
CHINESE (55)
ENGLISH (55)
MALAY (55)
{公司, 企业, 市场, 集团, 投资}
{business, companies, growth, market, financial}
{经济, 增长, 亚洲, 全球, 国际}
{kewangan, dana, pasaran, ekonomi, bank}
{pelaburan, penjualan, firma, taruhan, saham}
|< < > >|

Select a data source: DB1

Repository: ENGLISH
Cluster: E13
Search

Straits Times
United States
Monetary Authority of Singapore
Economic Development Board
Hong Kong
South Korea
joint venture
private sector
gross domestic product
interest rates

Translation(s)

CHINESE 中外合资 Search
MALAY usaha sama Search

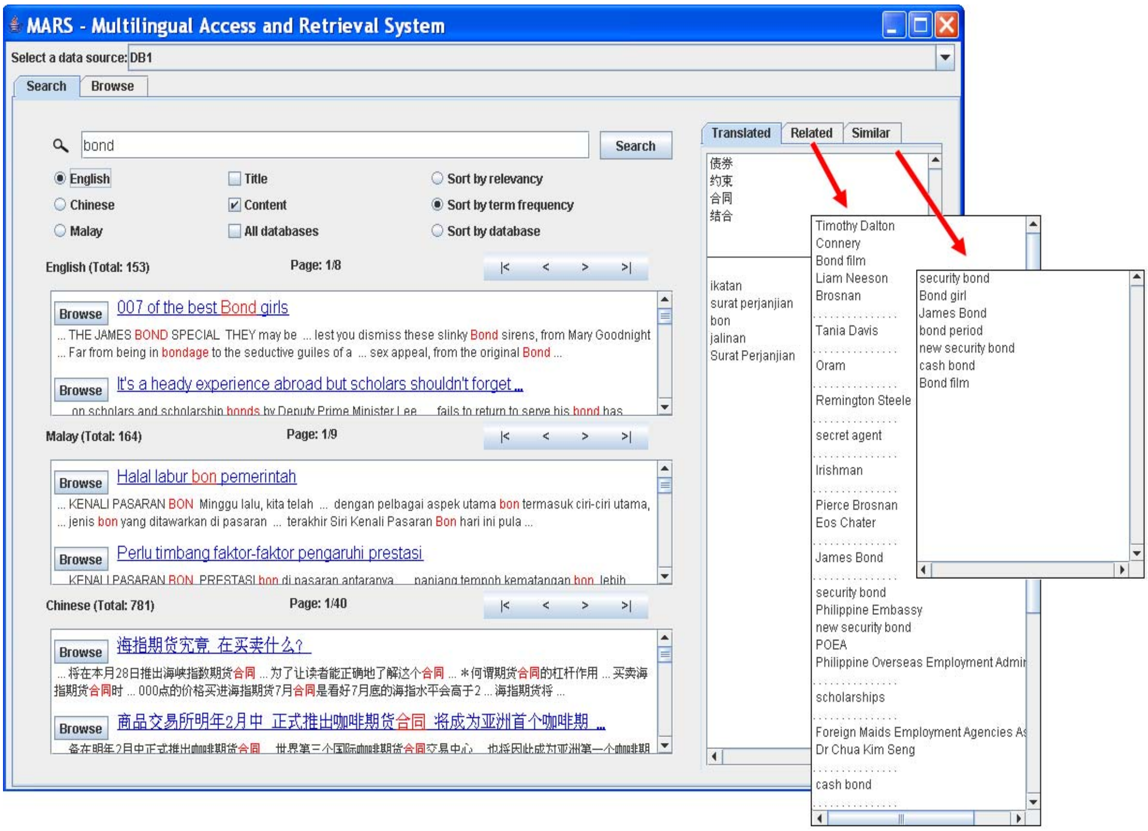
Documents (Total: 1264)

Page: 1/64

NTUC Income eyes Vietnam for insurance venture
... the materialisation of our joint venture," he said ... . Mr Tan was unable to ...

Group may sue Manila govt over 'breach' in shipyard sale
... Philyards is a 60:40 joint venture between the Philippines' ... Magsaysay ...

Fig. 1 Browse mode in the MARS System



Select a data source: DB1

bond Search

● English □ Title ○ Sort by relevancy
○ Chinese ☑ Content ● Sort by term frequency
○ Malay □ All databases ○ Sort by database

English (Total: 153) Page: 1/8

Browse 007 of the best Bond girls
... THE JAMES BOND SPECIAL THEY may be ... lest you dismiss these slinky Bond sirens, from Mary Goodnight ... Far from being in bondage to the seductive guiles of a ... sex appeal, from the original Bond ...

Browse It's a heady experience abroad but scholars shouldn't forget ...
... on scholars and scholarship bonds by Deputy Prime Minister Lee ... fails to return to serve his bond has

Malay (Total: 164) Page: 1/9

Browse Halal labur bon pemerintah
... KENALI PASARAN BON Minggu lalu, kita telah ... dengan pelbagai aspek utama bon termasuk ciri-ciri utama, ... jenis bon yang ditawarkan di pasaran ... terakhir Siri Kenali Pasaran Bon hari ini pula ...

Browse Perlu timbang faktor-faktor pengaruhi prestasi
... KENALI PASARAN BON PRESTASI bon di pasaran antaranya ... panjang tempoh kematangan bon lebih

Chinese (Total: 781) Page: 1/40

Browse 海指期货究竟 在买卖什么？
... 将在本月28日推出海峡指数期货合同 ... 为了让读者能正确地了解这个合同 ... ＊何谓期货合同的杠杆作用 ... 买卖海指期货合同时 ... 000点的价格买进海指期货7月合同是看好7月底的海指水平会高于2 ... 海指期货将 ...

Browse 商品交易所明年2月中 正式推出咖啡期货合同 将成为亚洲首个咖啡期 ...
... 备在明年2月中正式推出咖啡期货合同 ... 世界第三个国际咖啡期货合同交易中心 ... 也将因此成为亚洲第一个咖啡期

债券
约束
合同
结合

ikatan
surat perjanjian
bon
jalinan
Surat Perjanjian

Timothy Dalton
Connery
Bond film
Liam Neeson
Brosnan
....
Tania Davis
....
Oram
....
Remington Steele
....
secret agent
....
Irishman
....
Pierce Brosnan
Eos Chater
....
James Bond
....
security bond
Philippine Embassy
new security bond
POEA
Philippine Overseas Employment Admir
....
scholarships
....
Foreign Maids Employment Agencies As
Dr Chua Kim Seng
....
cash bond
....

security bond
Bond girl
James Bond
bond period
new security bond
cash bond
Bond film

Fig. 2 Search mode in the MARS System

The right hand side of the browse panel provides the detail information about the selected cluster using three sub-panels, i.e. top, middle and bottom. The top panel displays a list of extracted terms from the selected cluster. User may narrow down the list of interested terms by using the search-text column on top. By clicking on a term in the list, its translations in other languages, if any, will be displayed in the middle sub-panel and the document containing the term will be listed in the bottom sub-panel. The "Search" buttons next to the term translations provide a short-cut to jump to the search mode with the corresponding term translation being cut and pasted over. Last but not least, user may simply click on any document listed in the bottom sub-panel to read the content of the document and its aligned documents in a pop-up window.

4.2 Search Mode

Search mode provides a means for comprehensive information retrieval. Refer to the figure 2, user may enter query in any of the selected languages to search for documents in all languages. The main difference is that query translation is done via bilingual terms extracted via the term alignment technology discussed earlier. For each retrieved document, documents with similar content in the other languages are also provided to supplement the searched results. This enables documents which are potentially relevant to the users be retrieved as some of these retrieved documents may not contain the translated terms at all.

On top of the query translation, other information such as related terms and similar terms to the query are shown at the tab panel on the right. Related terms are terms that correlate statistically with the query term and they are arranged by cluster, separated by dotted line in the list. Similar terms are longer terms that contains the query term in itself. Both the related terms and the similar terms provide user additional hints and guides to improve further queries.

5 Conclusion

The MARS system is developed to enable user to better navigate and search information from multilingual comparable corpora in a user-friendly graphical user interface. Query translation and document retrieval is enhanced by utilizing the in-domain bilingual terminologies and document alignment acquired from the comparable corpora itself, without limited by dictionaries and keyword matching.

Currently, the system only support simple query. Future work will improve on this to allow more general query.

References

Ernesto William De Luca, and Andreas Nurnberger. 2006. *A Word Sense-Oriented User Interface for Interactive Multilingual Text Retrieval*, In Proceedings of the Workshop Information Retrieval, Hildesheim.

M. Ranieri, E. Pianta, and L. Bentivogli. 2004. *Browsing Multilingual Information with the MultiSemCor Web Interface*, In Proceedings of the LREC-2004 Workshop "The amazing utility of parallel and comparable corpora", Lisban, Portugal.

Fatiha Sadat, Masatoshi Yoshikawa, Shunsuke Uemura. 2003. *Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach*, In Proceedings of the 6th international workshop on Information Retrieval with Asian Languages, vol. 1: pp. 57-64.

Raghavendra Udupa, K. Saravanan, A. Kumaran, Jagadeesh Jagarlamudi. 2008. *Mining named entity transliteration equivalents from comparable corpora.* In Proceedings of the 17th ACM conference on Information and knowledge management.

Tao Tao, and ChengXiang Zhai. 2005. *Mining comparable bilingual text corpora for cross-language information integration.* In Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Jarvelin, Martti Juhola, Heikki Keskustalo. 2007. *Creating and exploiting a comparable corpus in cross-language information retrieval.* ACM Transactions on Information System (TOIS), vol. 25(1): Article No 4.

Thuy Vu, Aiti Aw, Min Zhang. 2008. *Term extraction through unithood and termhood unification.* In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08), Hyderabad, India.

Thuy Vu, Aiti Aw, Min Zhang. 2009. *Feature-based Method for Document Alignment in Comparable News Corpora.* In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), Athens, Greece.

Demonstration of Joshua: An Open Source Toolkit for Parsing-based Machine Translation[*]

Zhifei Li, Chris Callison-Burch, Chris Dyer[†], Juri Ganitkevitch[+], Sanjeev Khudanpur, Lane Schwartz[⋆], Wren N. G. Thornton, Jonathan Weese, and **Omar F. Zaidan**
Center for Language and Speech Processing, Johns Hopkins University
† Computational Linguistics and Information Processing Lab, University of Maryland
+ Human Language Technology and Pattern Recognition Group, RWTH Aachen University
⋆ Natural Language Processing Lab, University of Minnesota

Abstract

We describe **Joshua** (Li et al., 2009a)[1], an open source toolkit for statistical machine translation. Joshua implements all of the algorithms required for translation via synchronous context free grammars (SCFGs): chart-parsing, n-gram language model integration, beam- and cube-pruning, and k-best extraction. The toolkit also implements suffix-array grammar extraction and minimum error rate training. It uses parallel and distributed computing techniques for scalability. We also provide a demonstration outline for illustrating the toolkit's features to potential users, whether they be newcomers to the field or power users interested in extending the toolkit.

1 Introduction

Large scale parsing-based statistical machine translation (e.g., Chiang (2007), Quirk et al. (2005), Galley et al. (2006), and Liu et al. (2006)) has made remarkable progress in the last few years. However, most of the systems mentioned above employ tailor-made, dedicated software that is not open source. This results in a high barrier to entry for other researchers, and makes experiments difficult to duplicate and compare. In this paper, we describe **Joshua**, a Java-based general-purpose open source toolkit for parsing-based machine translation, serving the same role as Moses (Koehn et al., 2007) does for regular phrase-based machine translation.

2 Joshua Toolkit

When designing our toolkit, we applied general principles of software engineering to achieve three major goals: *Extensibility*, *end-to-end coherence*, and *scalability*.

Extensibility: Joshua's codebase consists of a separate Java `package` for each major aspect of functionality. This way, researchers can focus on a single `package` of their choosing. Fuurthermore, extensible components are defined by Java `interface`s to minimize unintended interactions and unseen dependencies, a common hindrance to extensibility in large projects. Where there is a clear point of departure for research, a basic implementation of each `interface` is provided as an `abstract class` to minimize work necessary for extensions.

End-to-end Cohesion: An MT pipeline consists of many diverse components, often designed by separate groups that have different file formats and interaction requirements. This leads to a large number of scripts for format conversion and to facilitate interaction between the components, resulting in untenable and non-portable projects, and hindering repeatability of experiments. Joshua, on the other hand, integrates the critical components of an MT pipeline seamlessly. Still, each component can be used as a stand-alone tool that does not rely on the rest of the toolkit.

Scalability: Joshua, especially the decoder, is scalable to large models and data sets. For example, the parsing and pruning algorithms are implemented with dynamic programming strategies and efficient data structures. We also utilize suffix-array grammar extraction, parallel/distributed decoding, and bloom filter language models.

Joshua offers state-of-the-art quality, having been ranked 4th out of 16 systems in the French-English task of the 2009 WMT evaluation, both in automatic (Table 1) and human evaluation.

[*] This research was supported in part by the Defense Advanced Research Projects Agency's GALE program under Contract No. HR0011-06-2-0001 and the National Science Foundation under grants No. 0713448 and 0840112. The views and findings are the authors' alone.

[1] Please cite Li et al. (2009a) if you use Joshua in your research, and **not** this demonstration description paper.

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 25–28,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP

System	BLEU-4
google	31.14
lium	26.89
dcu	26.86
joshua	**26.52**
uka	25.96
limsi	25.51
uedin	25.44
rwth	24.89
cmu-statxfer	23.65

Table 1: BLEU scores for top primary systems on the WMT-09 French-English Task from Callison-Burch et al. (2009), who also provide human evaluation results.

2.1 Joshua Toolkit Features

Here is a short description of Joshua's main features, described in more detail in Li et al. (2009a):

- **Training Corpus Sub-sampling:** We support inducing a grammar from a subset of the training data, that consists of sentences needed to translate a particular test set. To accomplish this, we make use of the method proposed by Kishore Papineni (personal communication), outlined in further detail in (Li et al., 2009a). The method achieves a 90% reduction in training corpus size while maintaining state-of-the-art performance.

- **Suffix-array Grammar Extraction:** Grammars extracted from large training corpora are often far too large to fit into available memory. Instead, we follow Callison-Burch et al. (2005) and Lopez (2007), and use a source language suffix array to extract only rules that will actually be used in translating a particular test set. Direct access to the suffix array is incorporated into the decoder, allowing rule extraction to be performed for each input sentence individually, but it can also be executed as a standalone pre-processing step.

- **Grammar formalism:** Our decoder assumes a probabilistic synchronous context-free grammar (SCFG). It handles SCFGs of the kind extracted by Hiero (Chiang, 2007), but is easily extensible to more general SCFGs (as in Galley et al. (2006)) and closely related formalisms like synchronous tree substitution grammars (Eisner, 2003).

- **Pruning:** We incorporate beam- and cube-pruning (Chiang, 2007) to make decoding feasible for large SCFGs.

- **k-best extraction:** Given a source sentence, the chart-parsing algorithm produces a *hypergraph* representing an exponential number of derivation hypotheses. We implement the extraction algorithm of Huang and Chiang (2005) to extract the k most likely derivations from the hypergraph.

- **Oracle Extraction:** Even within the large set of translations represented by a hypergraph, some desired translations (e.g. the references) may not be contained due to pruning or inherent modeling deficiency. We implement an efficient dynamic programming algorithm (Li and Khudanpur, 2009) for finding the *oracle translations*, which are most similar to the desired translations, as measured by a metric such as BLEU.

- **Parallel and distributed decoding:** We support *parallel decoding* and a *distributed language model* that exploit multi-core and multi-processor architectures and distributed computing (Li and Khudanpur, 2008).

- **Language Models:** We implement three local n-gram language models: a straightforward implementation of the n-gram scoring function in Java, capable of reading standard ARPA backoff n-gram models; a native code bridge that allows the decoder to use the SRILM toolkit to read and score n-grams[2]; and finally a Bloom Filter implementation following Talbot and Osborne (2007).

- **Minimum Error Rate Training:** Joshua's MERT module optimizes parameter weights so as to maximize performance on a development set as measured by an automatic evaluation metric, such as BLEU. The optimization consists of a series of line-optimizations using the efficient method of Och (2003). More details on the MERT method and the implementation can be found in Zaidan (2009).[3]

[2] The first implementation allows users to easily try the Joshua toolkit without installing SRILM. However, users should note that the basic Java LM implementation is not as scalable as the SRILM native bridge code.

[3] The module is also available as a standalone application, *Z-MERT*, that can be used with other MT systems.

- **Variational Decoding:** *spurious ambiguity* causes the probability of an output string among to be split among many derivations. The goodness of a string is measured by the total probability of its derivations, which means that finding the best output string is computationally intractable. The standard Viterbi approximation is based on the most probable derivation, but we also implement a variational approximation, which considers all the derivations but still allows tractable decoding (Li et al., 2009b).

3 Demonstration Outline

The purpose of the demonstration is 4-fold: 1) to give newcomers to the field of statistical machine translation an idea of the state-of-the-art; 2) to show actual, live, end-to-end operation of the system, highlighting its main components, targeting potential users; 3) to illustrate, through visual aids, the underlying algorithms, for those interested in the technical details; and 4) to explain how those components can be extended, for potential *power users* who want to be familiar with the code itself.

The first component of the demonstration will be an interactive user interface, where arbitrary user input in a source language is entered into a web form and then translated into a target language by the system. This component specifically targets newcomers to SMT, and demonstrates the current state of the art in the field. We will have trained multiple systems (for multiple language pairs), hosted on a remote server, which will be queried with the sample source sentences.

Potential users of the system would be interested in seeing an actual operation of the system, in a similar fashion to what they would observe on their own machines when using the toolkit. For this purpose, we will demonstrate three main modules of the toolkit: the rule extraction module, the MERT module, and the decoding module. Each module will have a separate terminal window executing it, hence demonstrating both the module's expected output as well as its speed of operation.

In addition to demonstrating the functionality of each module, we will also provide accompanying visual aids that illustrate the underlying algorithms and the technical operational details. We will provide visualization of the search graph and the 1-best derivation, which would illustrate the functionality of the decoder, as well as alternative translations for phrases of the source sentence, and where they were learned in the parallel corpus, illustrating the functionality of the grammar rule extraction. For the MERT module, we will provide figures that illustrate Och's efficient line search method.

4 Demonstration Requirements

The different components of the demonstration will be spread across at most 3 machines (Figure 1): one for the live "instant translation" user interface, one for demonstrating the different components of the system and algorithmic visualizations, and one designated for technical discussion of the code. We will provide the machines ourselves and ensure the proper software is installed and configured. However, we are requesting that large LCD monitors be made available, if possible, since that would allow more space to demonstrate the different components with clarity than our laptop displays would provide. We will also require Internet connectivity for the live demonstration, in order to gain access to remote servers where trained models will be hosted.

(Software and documentation at: `http://cs.jhu.edu/~ozaidan/zmert`.)

References

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of ACL*.

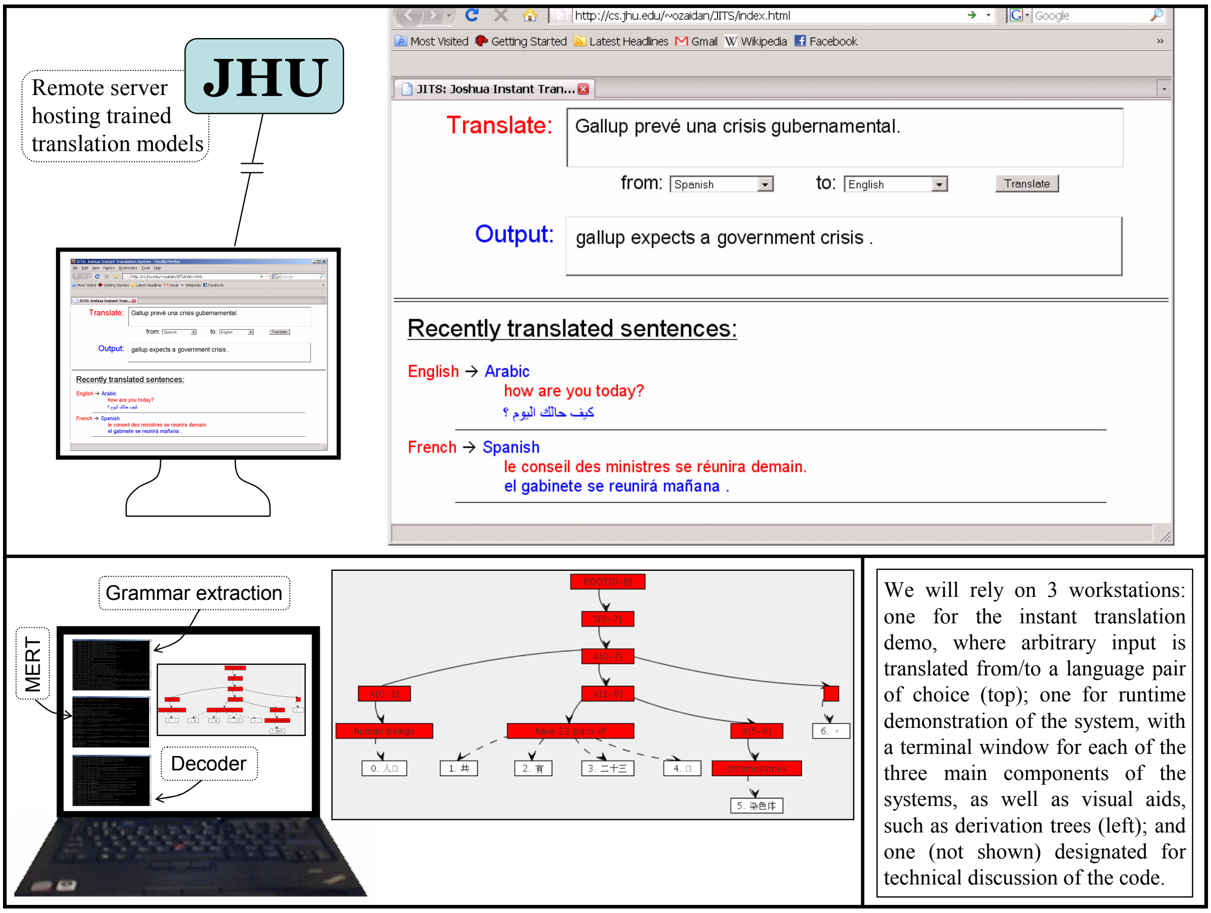
Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of ACL*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the ACL/Coling*.

Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proceedings of the International Workshop on Parsing Technologies*.



Translate: Gallup prevé una crisis gubernamental.

from: Spanish to: English

Output: gallup expects a government crisis .

Recently translated sentences:

English → Arabic
how are you today?
كيف حالك اليوم ؟

French → Spanish
le conseil des ministres se réunira demain.
el gabinete se reunirá mañana .

We will rely on 3 workstations: one for the instant translation demo, where arbitrary input is translated from/to a language pair of choice (top); one for runtime demonstration of the system, with a terminal window for each of the three main components of the systems, as well as visual aids, such as derivation trees (left); and one (not shown) designated for technical discussion of the code.

Figure 1: Proposed setup of our demonstration. When this paper is viewed as a PDF, the reader may zoom in further to see more details.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*.

Zhifei Li and Sanjeev Khudanpur. 2008. A scalable decoder for parsing-based machine translation with equivalent language model state maintenance. In *Proceedings Workshop on Syntax and Structure in Statistical Translation*.

Zhifei Li and Sanjeev Khudanpur. 2009. Efficient extraction of oracle-best translations from hypergraphs. In *Proceedings of NAACL*.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009a. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

Zhifei Li, Jason Eisner, and Sanjeev Khudanpur. 2009b. Variational decoding for statistical machine translation. In *Proceedings of ACL*.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment templates for statistical machine translation. In *Proceedings of the ACL/Coling*.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of EMNLP-CoLing*.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of ACL*.

David Talbot and Miles Osborne. 2007. Randomised language modelling for statistical machine translation. In *Proceedings of ACL*.

Omar F. Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

WikiBABEL: A Wiki-style Platform for Creation of Parallel Data

A Kumaran[†] **K Saravanan**[†] **Naren Datha**[*] **B Ashok**[*] **Vikram Dendi**[‡]

[†]Multilingual Systems Research
Microsoft Research India

[*]Advanced Development & Prototyping
Microsoft Research India

[‡]Machine Translation Incubation
Microsoft Research

Abstract

In this demo, we present a wiki-style platform – WikiBABEL – that enables easy collaborative creation of multilingual content in many non-English Wikipedias, by leveraging the relatively larger and more stable content in the English Wikipedia. The platform provides an intuitive user interface that maintains the user focus on the multilingual Wikipedia content creation, by engaging search tools for easy discoverability of related English source material, and a set of linguistic and collaborative tools to make the content translation simple. We present two different usage scenarios and discuss our experience in testing them with real users. Such integrated content creation platform in Wikipedia may yield as a by-product, parallel corpora that are critical for research in statistical machine translation systems in many languages of the world.

1 Introduction

Parallel corpora are critical for research in many natural language processing systems, especially, the Statistical Machine Translation (SMT) and Crosslingual Information Retrieval (CLIR) systems, as the state-of-the-art systems are based on statistical learning principles; a typical SMT system in a pair of language requires large parallel corpora, in the order of a few million parallel sentences. Parallel corpora are traditionally created by professionals (in most cases, for business or governmental needs) and are available only in a few languages of the world. The prohibitive cost associated with creating new parallel data implied that the SMT research was restricted to only a handful of languages of the world. To make such research possible widely, it is important that innovative and inexpensive ways of creating parallel corpora are found. Our research explores such an avenue: by involving the user community in creation of parallel data.

In this demo, we present a community collaboration platform – WikiBABEL – which enables the creation of multilingual content in Wikipedia. WikiBABEL leverages two significant facts with respect to Wikipedia data: First, there is a large skew between the content of English and non-English Wikipedias. Second, while the original content creation requires subject matter experts, subsequent translations may be effectively created by people who are fluent in English and the target language. In general, we do expect the large English Wikipedia to provide source material for multilingual Wikipedias; however on specific topics specific multilingual Wikipedia may provide the source material (*http://ja.wikipedia.org/wiki/俳句* may be better than *http://en.wikipedia.org/wiki/haiku*). We leverage these facts in the WikiBABEL framework, enabling a community of interested native speakers of a language, to create content in their respective language Wikipedias. We make such content creation easy by integrating linguistic tools and resources for translation, and collaborative mechanism for storing and sharing knowledge among the users. Such methodology is expected to generate comparable data (similar, but not the same content), from which parallel data may be mined subsequently (Munteanu et al, 2005) (Quirk et al, 2007).

We present here the WikiBABEL platform, and trace its evolution through two distinct usage versions: First, as a standalone deployment providing a community of users a translation platform on hosted Wikipedia data to generate parallel corpora, and second, as a transparent edit layer on top of Wikipedias to generate comparable corpora. Both paradigms were used for user testing, to gauge the usability of the tool and the viability of the approach for content creation in multilingual Wikipedias. We discuss the implementations and our experience with each of the above scenarios. Such experience may be very valuable in fine-tuning methodologies for community creation of various types of linguistic data. Community contributed efforts may perhaps be the only way to collect sufficient corpora effectively and economically, to enable research in many resource-poor languages of the world.

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 29–32,
Suntec, Singapore, 3 August 2009.

2 Architecture of WikiBABEL

The architecture of WikiBABEL is as illustrated in Figure 1: Central to the architecture is the *WikiBABEL* component that coordinates the interaction between its linguistic and collaboration components, and the users and the Wikipedia system. WikiBABEL architecture is designed to support a host of linguistic tools and resources that may be helpful in the content creation process: *Bilingual dictionaries* for providing for word-level translations, allowing user customization of domain-specific, or even, user-specific bilingual dictionaries. Also available are *machine translation and transliteration* systems for rough initial translation [or transliteration] of a source language string at sentential/phrasal levels [or names] to the intended target language. As the quality of automatic translations are rarely close to human quality translations, the user may need to correct any such automatically translated or transliterated content, and an intuitive edit framework provides tools for such corrections. A *collaborative translation memory* component stores all the user corrections (or, sometimes, their selection from a set of alternatives) of machine translations, and makes them available to the community as a translation help ('*tribe knowledge*'). Voting mechanisms are available that may prioritize more frequently chosen alternatives as preferred suggestions for subsequent users. The *user-management* tracks the user demographic information, and their contributions (its quality and quantity) for possible recognition. The user interface features are implemented as light-weight components, requiring minimal server-side interaction. Finally, the architecture is designed open, to integrate any user-developed tools and resources easily.

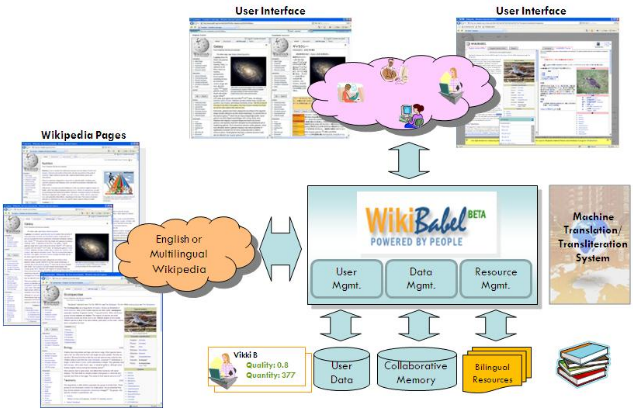


Figure 1: WikiBABEL Architecture

3 WikiBABEL on Wikipedia

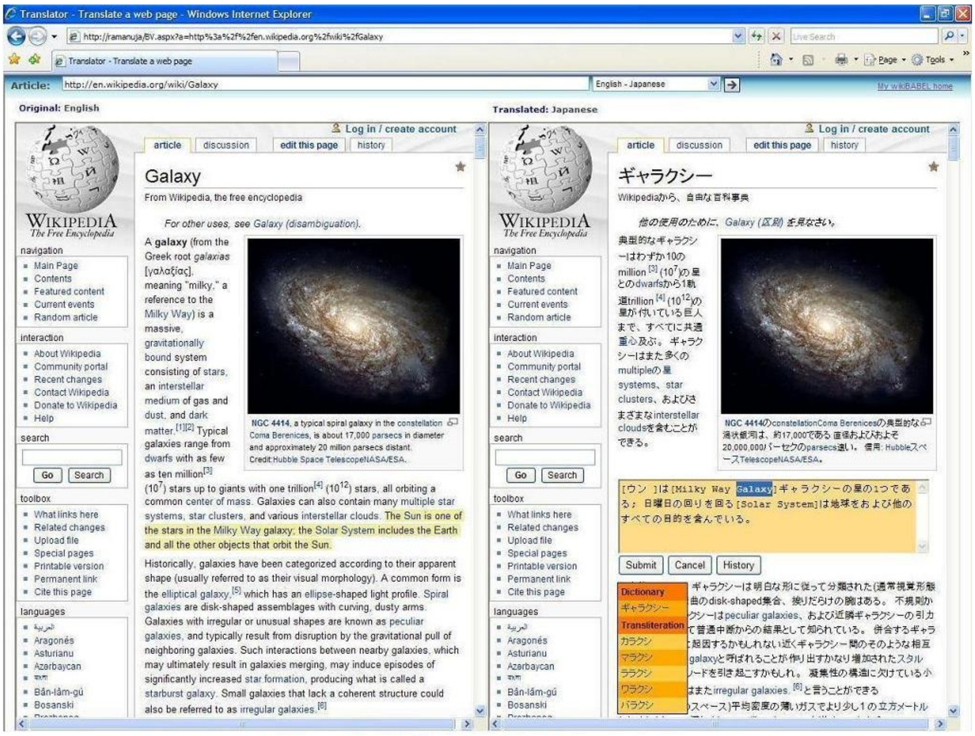
IN this section we discuss Wikipedia content and user characteristics and outline our experience with the two versions on Wikipedia.

3.1 Wikipedia: User & Data Characteristics

Wikipedia content is acknowledged to be on par with the best of the professionally created resources (Giles, 2005) and is used regularly as academic reference (Rainie *et al*., 2007). However, there is a large disparity in content between English and other language Wikipedias. English Wikipedia - the largest - has about 3.5 Million topics, but with an exception of a dozen or so Western European and East Asian languages, most of the 250-odd languages have less than 1% of English Wikipedia content (Wikipedia, 2009). Such skew, despite the size of the respective user population, indicates a large room for growth in many multilingual Wikipedias. On the contribution side, Wikipedia has about 200,000 contributors (> 10 total contributions); but only about 4% of them are very active (> 100 contributions per month). The general perception that a few very active users contributed to the bulk of Wikipedia was disputed in a study (Swartz, 2006) that claims that large fraction of the content were created by those who made very few or occasional contributions that are primarily editorial in nature. It is our strategy to provide a platform for easy multilingual Wikipedia content creation that may be harvested for parallel data.

3.2 Version 1: A Hosted Portal

In our first version, a set of English Wikipedia topics (stable non-controversial articles, typically from Medicine, Healthcare, Science & Technology, Literature, etc.) were chosen and hosted in our WikiBABEL portal. Such set of articles is already available as *Featured Articles* in most Wikipedias. English Wikipedia has a set of ~1500 articles that are voted by the community as stable and well written, spanning many domains, such as, Literature, Philosophy, History, Science, Art, etc. The user can choose any of these Wikipedia topics to translate to the target language and correct the machine translation errors. Once a topic is chosen, a two-pane window is presented to the user, as shown in Figure 2, in which the original English Wikipedia article is shown in the left panel and a rough translation of the same article in the user-chosen target language is presented in the right panel. The right panel has the same look and feel as the original



Galaxy

From Wikipedia, the free encyclopedia

For other uses, see Galaxy (disambiguation).

A **galaxy** (from the Greek root *galaxias* [γαλαξίας], meaning "milky," a reference to the Milky Way) is a massive, gravitationally bound system consisting of stars, an interstellar medium of gas and dust, and dark matter.[1][2] Typical galaxies range from dwarfs with as few as ten million[3] (10^7) stars up to giants with one trillion[4] (10^{12}) stars, all orbiting a common center of mass. Galaxies can also contain many multiple star systems, star clusters, and various interstellar clouds. The Sun is one of the stars in the Milky Way galaxy; the Solar System includes the Earth and all the other objects that orbit the Sun.

NGC 4414, a typical spiral galaxy in the constellation Coma Berenices, is about 17,000 parsecs in diameter and approximately 20 million parsecs distant. Credit:Hubble Space TelescopeNASA/ESA.

Historically, galaxies have been categorized according to their apparent shape (usually referred to as their visual morphology). A common form is the elliptical galaxy,[5] which has an ellipse-shaped light profile. Spiral galaxies are disk-shaped assemblages with curving, dusty arms. Galaxies with irregular or unusual shapes are known as peculiar galaxies, and typically result from disruption by the gravitational pull of neighboring galaxies. Such interactions between nearby galaxies, which may ultimately result in galaxies merging, may induce episodes of significantly increased star formation, producing what is called a starburst galaxy. Small galaxies that lack a coherent structure could also be referred to as irregular galaxies.[6]

ギャラクシー

Wikipediaから、自由な百科事典

他の使用のために、Galaxy (区別) を見なさい。

典型的なギャラクシーはわずか10の million [3] (10^7)の星とのdwarfsから1軌道trillion [4] (10^{12})の星が付いている巨人まで、すべてに共通重心及ぶ。ギャラクシーはまた多くのmultipleの星systems、star clusters、およびさまざまなinterstellar cloudsを含むことができる。

NGC 4414のconstellationComa Berenicesの典型的な渦状銀河は、約17,000である直径およびおよそ20,000,000パーセクのparsecs遠い。信用: HubbleスペースTelescopeNASA/ESA。

[ウン]は[Milky Way Galaxy]ギャラクシーの星の1つである; 日曜日の回りを回る[Solar System]は地球をおよび他のすべての目的を含んでいる。

ギャラクシーは明白な形に従って分類された(通常視覚形態
曲のdisk-shaped集合、挨りだらけの腕はある。不規則か
クシーはpeculiar galaxies、および近隣ギャラクシーの引力
て普通中断からの結果として知られている。併合するギャラ
起因するかもしれない近くギャラクシー間のそのような相互
galaxyと呼ばれることが作り出すかなり増加されたスタル
ードを引き起こすかもしれ。凝集性の構造に欠けている小
はまたirregular galaxies.[6]と言うことができる
スペース)平均密度の薄いガスでより少し1の立方メートル

Figure 2: WikiBABEL Version 1

English Wikipedia article, and is editable, while the left panel is primarily intended for providing source material for reference and context, for the translation correction. On mouse-over the parallel sentences are highlighted, linking visually the related text on both panels. On a mouse-click, an edit-box is opened *in-place* in the right panel, and the current content may be edited. As mentioned earlier, integrated linguistic tools and resources may be invoked during edit process, to help the user. Once the article reaches sufficient quality as judged by the users, the content may be transferred to target language Wikipedia, effectively creating a new topic in the target language Wikipedia.

User Feedback: We field tested our first version with a set of Wikipedia users, and a host of amateur and professional translators. The primary feedback we got was that such efforts to create content in multilingual Wikipedia was well appreciated. The testing provided much quantitative (in terms of translation time, effort, etc.) and qualitative (user experience) measures and feedback. The details are available in (Kumaran et al., 2008), and here we provide highlights only:

- Integrated linguistic resources (e.g., bilingual dictionaries, transliteration systems, etc.) were appreciated by all users.

- Amateur users used the automatic translations (in direct correlation with its quality), and improved their throughput up to 40%.

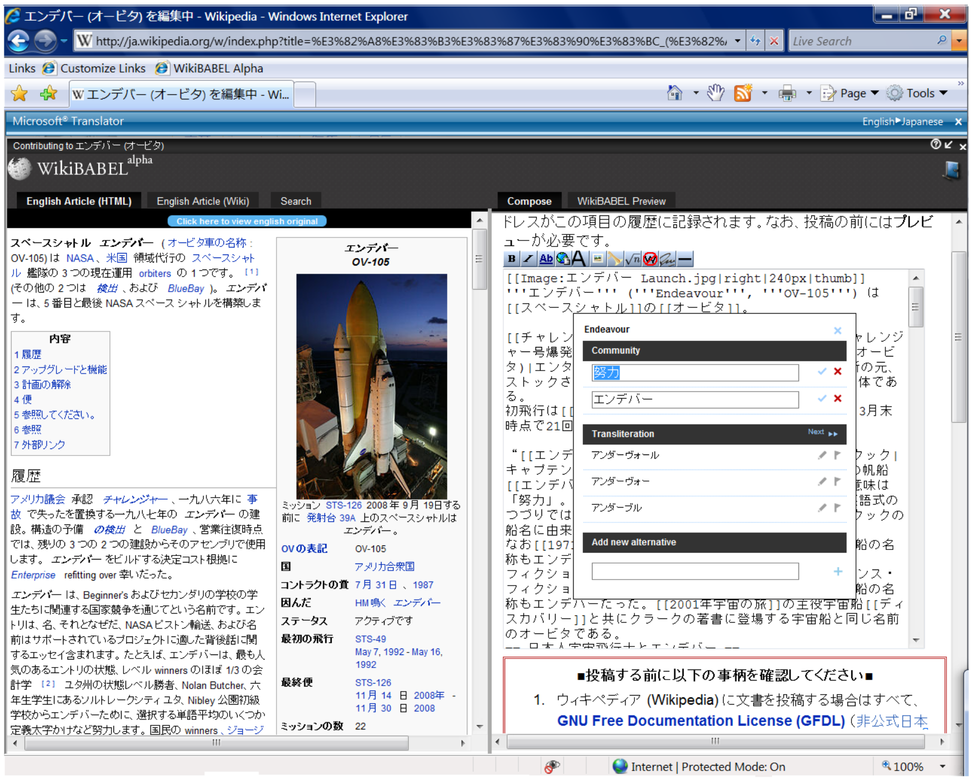
- In contrast, those who were very fluent in both the languages were distracted by the quality of translations, and were slowed by 30%. In most cases, they preferred to redo the entire translations, rather than considering and correcting the rough translation.

- One qualitative feedback from the Wikipedia community is that the sentence-by-sentence translation enforced by the portal is not in tune with their philosophy of user-decided content for the target topic.

We used the feedback from the version 1, to re-design WikiBABEL in version 2.

3.3 Version 2: As a Transparent Edit Layer

In our second version, we implemented the significant feedback from Wikipedians, pertaining to source content selection and the user contribution. In this version, we delivered the WikiBABEL experience as an add-on to Wikipedia, as a semi-transparent overlay that augments the basic Wikipedia edit capabilities without taking the contributor away from the site. Capable of being launched with one click (via a bookmarklet, or a browser plug-in, or as a potential server side integration with Wikipedia), the new version offered a more seamless workflow and integrated linguistic and collaborative components. This add-on may be invoked on Wikipedia itself, providing all WikiBABEL functionalities. In a typical WikiBABEL usage scenario, a Wikipedia



エンデバー (オービタ) を編集中 - Wikipedia - Windows Internet Explorer
http://ja.wikipedia.org/w/index.php?title=%E3%82%A8%E3%83%B3%E3%83%87%E3%83%90%E3%83%BC_(%E3%82%
Live Search
Links
Customize Links
WikiBABEL Alpha
エンデバー (オービタ) を編集中 - Wi...
Page
Tools
Microsoft® Translator
English▸Japanese
Contributing to エンデバー (オービタ)
WikiBABEL alpha
English Article (HTML)
English Article (Wiki)
Search
Click here to view english original
Compose
WikiBABEL Preview
Internet | Protected Mode: On
100%

スペースシャトル エンデバー (オービタ車の名称 : OV-105) は NASA 、米国 領域代行の スペースシャトル 艦隊の 3 つの現在運用 orbiters の 1 つです。 [1] (その他の 2 つは 検出 、および BlueBay)。 エンデバー は、5 番目と最後 NASA スペース シャトルを構築します。

内容
1 履歴
2 アップグレードと機能
3 計画の解除
4 便
5 参照してください。
6 参照
7 外部リンク

履歴

アメリカ議会 承認 チャレンジャー 、一九八六年に 事故 で失ったを置換する一九八七年の エンデバー の建設。構造の予備 の検出 と BlueBay 、営業往復時点では、残りの 3 つの 2 つの建設からそのアセンブリで使用します。 エンデバー をビルドする決定コスト根拠に Enterprise refitting over 幸いだった。

エンデバー は、Beginner's およびセカンダリの学校の学生たちに関連する国家競争を通じてという名前です。エントリは、名、それとなぜだ、NASA ピストン輸送、および名前はサポートされているプロジェクトに適した背後話に関するエッセイ含まれます。たとえば、エンデバーは、最も人気のあるエントリの状態、レベル winners のほぼ 1/3 の会計学 [2] ユタ州の状態レベル勝者、Nolan Butcher、六年生学生にあるソルトレークシティ ユタ、Nibley 公園初級学校からエンデバーために、選択する単語平均のいくつか定義太字かけなど努力します。国民の winners 、ジョージ

エンデバー
OV-105

ミッション STS-126 2008 年 9 月 19日する前に 発射台 39A 上のスペースシャトルは エンデバー 。

OV の表記	OV-105
国	アメリカ合衆国
コントラクトの貫	7 月 31 日 、 1987
因んだ	HM 鳴く エンデバー
ステータス	アクティブです
最初の飛行	STS-49 May 7, 1992 - May 16, 1992
最終便	STS-126 11 月 14 日 2008年 - 11 月 30 日 2008
ミッションの数	22

ドレスがこの項目の履歴に記録されます。なお、投稿の前にはプレビューが必要です。

```
[[Image:エンデバー Launch.jpg|right|240px|thumb]]
'''エンデバー''' ('''Endeavour''', '''OV-105''') は
[[スペースシャトル]]の[[オービタ]]。
```

Endeavour

Community

努力 ✓ ✗

エンデバー ✓ ✗

Transliteration Next ▸▸

アンダーヴォール

アンダーヴォー

アンダーブル

Add new alternative

```
称もエンデバーだった。[[2001年宇宙の旅]]の主役宇宙船[[ディ
スカバリー]]と共にクラークの著書に登場する宇宙船と同じ名前
のオービタである。
```

■投稿する前に以下の事柄を確認してください■

1. ウィキペディア (Wikipedia) に文書を投稿する場合はすべて、 GNU Free Documentation License (GFDL) (非公式日本

Figure 3: WikiBABEL Version 2

content creator may be at an English Wikipedia article for which no corresponding article exists in the target language, or at target language Wikipedia article which has much less content compared to the corresponding English article.

The WikiBABEL user interface in this version is as shown in Figure 3. The source English Wikipedia article is shown in the left panel tabs, and may be toggled between English and the target language; also it may be viewed in HTML or in Wiki-markup. The right panel shows the target language Wikipedia article (if it exists), or a newly created stub (otherwise); either case, the right panel presents a *native target language Wikipedia edit page*, for the chosen topic. The left panel content is used as a reference for content creation in target language Wikipedia in the right panel. The user may compose the target language Wikipedia article, either by dragging-and-dropping translated content from the left to the right panel (into the target language Wikipedia editor), or add new content as a typical Wikipedia user would. To enable the user to stay within WikiBABEL for their content research, we have provided the capability to search through other Wikipedia articles in the left panel. All linguistic and collaborative features are available to the users in the right panel, as in the previous version. The default target language Wikipedia preview is at any time. While the user testing of this implementation is still in the preliminary stages, we wish to make the following observations on the methodology:

- There is a marked shift of focus from "*translation from English Wikipedia article*" to "*content creation in target Wikipedia*".
- The user is never taken away from Wikipedia site, requiring optionally only Wikipedia credentials. The content is created directly in the target Wikipedia.

The WikiBABEL Version 2 prototype will be made available externally in the future.

References

Kumaran, A, Saravanan, K and Maurice, S. WikiBABEL: Community Creation of Multilingual Data. *WikiSYM 2008 Conference*, 2008.

Munteanu, D. and Marcu, D. Improving the MT performance by exploiting non-parallel corpora. *Computational Linguistics*. 2005.

Giles, J. Internet encyclopaedias go head to head. Nature. 2005. *doi:10.1038/438900a*.

Quirk, C., Udupa, R. U. and Menezes, A. Generative models of noisy translations with app. to parallel fragment extraction. *MT Summit XI*, 2007.

Rainie, L. and Tancer, B. Pew Internet and American Life. *http://www.pewinternet.org/*.

Swartz, A. Raw thought: Who writes Wikipedia? 2006. *http://www.aaronsw.com/*.

Wikipedia Statistics, 2009.*http://stats.wikimedia.org/*.

System for Querying Syntactically Annotated Corpora

Petr Pajas
Charles Univ. in Prague, MFF ÚFAL
Malostranské nám. 25
118 00 Prague 1 – Czech Rep.
pajas@ufal.mff.cuni.cz

Jan Štěpánek
Charles Univ. in Prague, MFF ÚFAL
Malostranské nám. 25
118 00 Prague 1 – Czech Rep.
stepanek@ufal.mff.cuni.cz

Abstract

This paper presents a system for querying treebanks. The system consists of a powerful query language with natural support for cross-layer queries, a client interface with a graphical query builder and visualizer of the results, a command-line client interface, and two substitutable query engines: a very efficient engine using a relational database (suitable for large static data), and a slower, but paralel-computing enabled, engine operating on treebank files (suitable for "live" data).

1 Introduction

Syntactically annotated treebanks are a great resource of linguistic information that is available hardly or not at all in flat text corpora. Retrieving this information requires specialized tools. Some of the best-known tools for querying treebanks include TigerSEARCH (Lezius, 2002), TGrep2 (Rohde, 2001), MonaSearch (Maryns and Kepser, 2009), and NetGraph (Mírovský, 2006). All these tools dispose of great power when querying a single annotation layer with nodes labeled by "flat" feature records.

However, most of the existing systems are little equipped for applications on structurally complex treebanks, involving for example multiple interconnected annotation layers, multi-lingual parallel annotations with node-to-node alignments, or annotations where nodes are labeled by attributes with complex values such as lists or nested attribute-value structures. The Prague Dependency Treebank 2.0 (Hajič and others, 2006), PDT 2.0 for short, is a good example of a treebank with multiple annotation layers and richly-structured attribute values. NetGraph was a tool traditionally used for querying over PDT, but still it does not directly support cross-layer queries, unless the layers are merged together at the cost of loosing some structural information.

The presented system attempts to combine and extend features of the existing query tools and resolve the limitations mentioned above. We are grateful to an anonymous referee for pointing us to ANNIS2 (Zeldes and others, 2009) – another system that targets annotation on multiple levels.

2 System Overview

Our system, named PML Tree Query (PML-TQ), consists of three main components (discussed further in the following sections):

- an expressive *query language* supporting cross-layer queries, arbitrary boolean combinations of statements, able to query complex data structures. It also includes a sublanguage for generating listings and non-trivial statistical reports, which goes far beyond statistical features of e.g. TigerSearch.
- client interfaces: a graphical user interface with a graphical query builder, a customizable visualization of the results and a command-line interface.
- two interchangeable engines that evaluate queries: a very efficient engine that requires the treebank to be converted into a relational database, and a somewhat slower engine which operates directly on treebank files and is useful especially for data in the process of annotation and experimental data.

The query language applies to a generic data model associated with an XML-based data format called Prague Markup Language or PML (Pajas and Štěpánek, 2006). Although PML was developed in connection with PDT 2.0, it was designed as a universally applicable data format based on abstract data types, completely independent of a

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 33–36,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP

particular annotation schema. It can capture simple linear annotations as well as annotations with one or more richly structured interconnected annotation layers. A concrete PML-based format for a specific annotation is defined by describing the data layout and XML vocabulary in a special file called PML Schema and referring to this schema file from individual data files.

It is relatively easy to convert data from other formats to PML without loss of information. In fact, PML-TQ is implemented within the TrEd framework (Pajas and Štěpánek, 2008), which uses PML as its native data format and already offers all kinds of tools for work with treebanks in several formats using on-the-fly transformation to PML (for XML input via XSLT).

The whole framework is covered by an open-source license and runs on most current platforms. It is also language and script independent (operating internally with Unicode).

The graphical client for PML-TQ is an extension to the tree editor TrEd that already serves as the main annotation tool for treebank projects (including PDT 2.0) in various countries. The client and server communicate over the HTTP protocol, which makes it possible to easily use PML-TQ engine as a service for other applications.

3 Query Language

A PML-TQ query consists of a part that selects nodes in the treebank, and an optional part that generates a report from the selected occurrences.

The selective part of the query specifies conditions that a group of nodes must satisfy to match the query. The conditions can be formulated as arbitrary boolean combinations of subqueries and simple statements that can express all kinds of relations between nodes and/or attribute values. This part of the query can be visualized as a graph with vertices representing the matching nodes, connected by various types of edges. The edges (visualized by arrows of different colors and styles) represent various types of relations between the nodes. There are four kinds of these relations:

- topological relations (*child, descendant depth-first-precedes, order-precedes, same-tree-as, same-document-as*) and their reversed counterparts (*parent, ancestor, depth-first-follows, order-follows*)
- inter- or cross-layer ID-based references
- user-implemented relations, i.e. relations whose low-level implementation is provided by the user as an extension to PML-TQ[1] (for example, we define relations *eparent* and *echild* for PDT 2.0 to distinguish effective dependency from technical dependency).
- transitive closures of the preceding two types of relations (e.g. if `coref_text.rf` is a relation representing textual coreference, then `coref_text.rf{4,}` is a relation representing chains of textual coreference of length at least 4).

The query can be accompanied by an optional part consisting of a chain of output filters that can be used to extract data from the matching nodes, compute statistics, and/or format and post-process the results of a query.

Let us examine these features on an example of a query over PDT 2.0, which looks for Czech words that have a patient or effect argument in infinitive form:

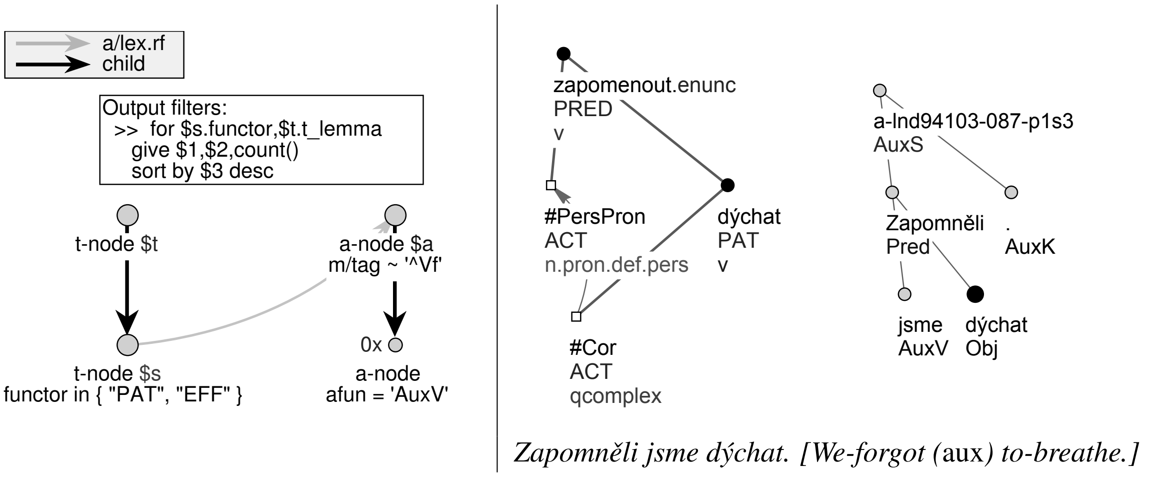
```
t-node $t := [
  child t-node $s := [
    functor in { "PAT", "EFF" },
    a/lex.rf $a ] ];
a-node $a := [
  m/tag ~ '^Vf',
  0x child a-node [ afun = 'AuxV' ] ];

>> for $s.functor,$t.t_lemma
   give $1, $2, count()
   sort by $3 desc
```

The square brackets enclose conditions regarding one node, so `t-node $t := [...]` is read *'t-node $t with ... '*. Comma is synonymous with logical `and`. See Fig. 3 for the graphical representation of the query and one match.

This particular query selects occurrences of a group of three nodes, `$t`, `$s`, and `$a` with the following properties: `$t` and `$s` are both of type `t-node`, i.e. nodes from a tectogrammatical tree (the types are defined in the PML Schema for the PDT 2.0); `$s` is a child of `$t`; the `functor` attribute of `$s` has either the value `PAT` or `EFF`; the node `$s` points to a node of type `a-node`, named `$a`, via an ID-based reference `a/lex.rf` (this expression in fact retrieves value of an attribute `lex.rf` from an attribute-value structure stored in the attribute `a` of `$s`); `$a` has an attribute `m` carrying an attribute-value structure with the attribute

[1] In some future version, the users will also be able to define new relations as separate PML-TQ queries.



Output filters:
>> for $s.functor,$t.t_lemma
give $1,$2,count()
sort by $3 desc

Zapomněli jsme dýchat. [We-forgot (aux) *to-breathe.]*

Figure 1: Graphical representation of a query (left) and a result spanning two annotation layers

`tag` matching regular expression `^Vf` (in PDT 2.0 tag set this indicates that `$a` is an infinitive); `$a` has no child node that is an auxiliary verb (`afun = 'AuxV'`). This last condition is expressed as a sub-query with zero occurrences (`0x`).

The selective part of the query is followed by one output filter (starting with `>>`). It returns three values for each match: the functor of `$s`, the tectogrammatical lemma of `$t`, and for each distinct pair of these two values the number of occurrences of this pair counted over the whole matching set. The output is ordered by the 3rd column in the descending order. It may look like this:

```
PAT     možnost        115
PAT     schopný        110
EFF     a               85
PAT     #Comma          83
PAT     rozhodnout_se   75
```

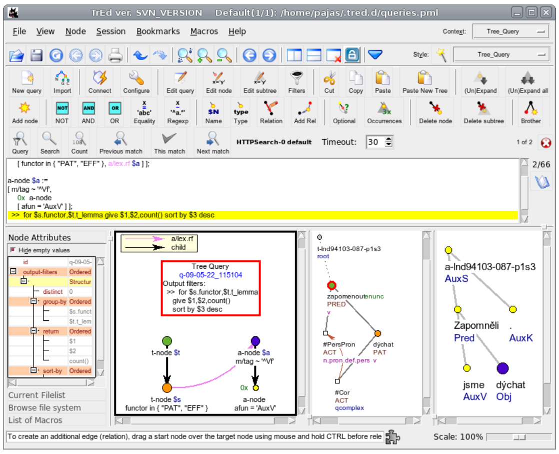
In the PML data model, attributes (like `a` of `$t`, `m` of `$a` in our example) can carry complex values: attribute-value structures, lists, sequences of named elements, which in turn may contain other complex values. PML-TQ addresses values nested within complex data types by attribute paths whose notation is somewhat similar to XPath (e.g. `m/tag` or `a/[2]/aux.rf`). An attribute path evaluated on a given node may return more than one value. This happens for example when there is a list value on the attribute path: the expression `m/w/token='a'` where `m` is a list of attribute-value structures reads as *some one value returned by* `m/w/token` *equals 'a'*. By prefixing the path with a $\star$, we may write *all values returned by* `m/w/token` *equal 'a'* as `*m/w/token='a'`.

We can also fix one value returned by an attribute path using the `member` keyword and query it the same way we query a node in the treebank:

```
t-node $n:= [
  member bridging [
    type = "CONTRAST",
    target.rf t-node [ functor="PAT" ]]]
```

where `bridging` is an attribute of `t-node` containing a list of labeled graph edges (attribute-value structures). We select one that has type `CONTRAST` and points to a node with functor PAT.

4 Query Editor and Client



```
  [ functor in { "PAT", "EFF" }, a/lex.rf $a ] ];

a-node $a :=
[ m/tag ~ '^Vf',
  0x  a-node
  [ afun = 'AuxV' ] ];
>>  for $s.functor,$t.t_lemma give $1,$2,count() sort by $3 desc
```

id	q-09-05-
output-filters	Ordered
	Structur
distinct	0
group-by	Ordered
	$s.funct
	$t.t_lem
return	Ordered
	$1
	$2
	count()
sort-by	Ordered

Tree Query
q-09-05-22_115104
Output filters:
>> for $s.functor,$t.t_lemma
give $1,$2,count()
sort by $3 desc

To create an additional edge (relation), drag a start node over the target node using mouse and hold CTRL before rele

Figure 2: The PML-TQ graphical client in TrEd

The graphical user interface lets the user to build the query graphically or in the text form; in both cases it assists the user by offering available node-types, applicable relations, attribute paths, and values for enumerated data types. It communicates with the query engine and displays the results (matches, reports, number of occurrences).

Colors are used to indicate which node in the query graph corresponds to which node in the result. Matches from different annotation layers are displayed in parallel windows. For each result, the user can browse the complete document for context. Individual results can be saved in the PML format or printed to PostScript, PDF, or SVG. The user can also bookmark any tree from the result set, using the bookmarking features of TrEd. The queries are stored in a local file.[2]

5 Engines

For practical reasons, we have developed two engines that evaluate PML-TQ queries:

The first one is based on a translator of PML-TQ to SQL. It utilizes the power of modern relational databases[3] and provides excellent performance and scalability (answering typical queries over a 1-million-word treebank in a few seconds). To use this engine, the treebank must be, similarly to (Bird and others, 2006), converted into read-only database tables, which makes this engine more suitable for data that do not change too often (e.g. final versions of treebanks).

For querying over working data or data not likely to be queried repeatedly, we have developed an index-less query evaluator written in Perl, which performs searches over arbitrary data files sequentially. Although generally slower than the database implementation (partly due to the cost of parsing the input PML data format), its performance can be boosted up using a built-in support for parallel execution on a computer cluster.

Both engines are accessible through the identical client interface. Thus, users can run the same query over a treebank stored in a database as well as their local files of the same type.

When implementing the system, we periodically verify that both engines produce the same results on a large set of test queries. This testing proved invaluable not only for maintaining consistency, but also for discovering bugs in the two implementations and also for performance tuning.

6 Conclusion

We have presented a powerful open-source system for querying treebanks extending an established framework. The current version of the system is available at `http://ufal.mff.cuni.cz/~pajas/pmltq`.

Acknowledgments

This paper as well as the development of the system is supported by the grant Information Society of GA AV ČR under contract 1ET101120503 and by the grant GAUK No. 22908.

References

Steven Bird et al. 2006. Designing and evaluating an XPath dialect for linguistic queries. In *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering*, page 52. IEEE Computer Society.

Jan Hajič et al. 2006. The Prague Dependency Treebank 2.0. CD-ROM. Linguistic Data Consortium (CAT: LDC2006T01).

Wolfgang Lezius. 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis, IMS, University of Stuttgart, December. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4.

Hendrik Maryns and Stephan Kepser. 2009. Monasearch – querying linguistic treebanks with monadic second-order logic. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT 2009)*.

Jiří Mírovský. 2006. Netgraph: A tool for searching in Prague Dependency Treebank 2.0. In *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 211–222.

Petr Pajas and Jan Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *The 22nd International Conference on Computational Linguistics - Proceedings of the Conference*, volume 2, pages 673–680. The Coling 2008 Organizing Committee.

Petr Pajas and Jan Štěpánek. 2006. XML-based representation of multi-layered annotation in the PDT 2.0. In *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47.

Douglas L.T. Rohde. 2001. TGrep2 the next-generation search engine for parse trees. `http://tedlab.mit.edu/~dr/Tgrep2/`.

Amir Zeldes et al. 2009. Information structure in african languages: Corpora and tools. In *Proceedings of the Workshop on Language Technologies for African Languages (AFLAT), 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), Athens, Greece*, pages 17–24.

[2] The possibility of storing the queries in a user account on the server is planned.

[3] The system supports Oracle Database (version 10g or newer, the free XE edition is sufficient) and PostgreSQL (version at least 8.4 is required for complete functionality).

A NLG-based Application for Walking Directions

Michael Roth and Anette Frank
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{mroth,frank}@cl.uni-heidelberg.de

Abstract

This work describes an online application that uses Natural Language Generation (NLG) methods to generate walking directions in combination with dynamic 2D visualisation. We make use of third party resources, which provide for a given query (geographic) routes and landmarks along the way. We present a statistical model that can be used for generating natural language directions. This model is trained on a corpus of walking directions annotated with POS, grammatical information, frame-semantics and markup for temporal structure.

1 Introduction

The purpose of route directions is to inform a person, who is typically not familiar with his current environment, of how to get to a designated goal. Generating such directions poses difficulties on various conceptual levels such as the planning of the route, the selection of landmarks along the way (i.e. easily recognizable buildings or structures) and generating the actual instructions of how to navigate along the route using the selected landmarks as reference points.

As pointed out by Tom & Denis (2003), the use of landmarks in route directions allows for more effective way-finding than directions relying solely on street names and distance measures. An experiment performed in Tom & Denis' work also showed that people tend to use landmarks rather than street names when producing route directions themselves.

The application presented here is an early research prototype that takes a data-driven generation approach, making use of annotated corpora collected in a way-finding study. In contrast to previously developed NLG systems in this area (e.g. Dale et. al, 2002), one of our key features is the integration of a number of online resources to compute routes and to find salient landmarks. The information acquired from these resources can then be used to generate natural directions that are both easier to memorise and easier to follow than directions given by a classic route planner or navigation system.

The remainder of this paper is structured as follows: In Section 2 we introduce our system and describe the resources and their integration in the architecture. Section 3 describes our corpus-based generation approach, with Section 4 outlining our integration of text generation and visualisation. Finally, Section 5 gives a short conclusion and discusses future work.

2 Combining Resources

The route planner used in our system is provided by the *Google Maps API*[1]. Given a route computed in Google Maps, our system queries a number of online resources to determine landmarks that are adjacent to this route. At the time of writing, these resources are: *OpenStreetMaps*[2] for public transportation, the Wikipedia *WikiProject Geographical coordinates*[3] for salient buildings, statues and other objects, *Google AJAX Search API*[4] for "yellow pages landmarks" such as hotels and restaurants, and *Wikimapia*[5] for squares and other prominent places.

All of the above mentioned resources can be queried for landmarks either by a single GPS

[1] http://code.google.com/apis/maps/
[2] http://www.openstreetmap.org
[3] http://en.wikipedia.org/wiki/Wikipedia:WikiProject Geographical_coordinates
[4] http://code.google.com/apis/ajaxsearch
[5] http://www.wikimapia.org

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 37–40,
Suntec, Singapore, 3 August 2009.

coordinate (using the `LocalSearch` method in Google AJAX Search and web tools in Wikipedia) or an area of GPS coordinates (using URL based queries in Wikimapia and OpenStreetMaps). The following list describes the data formats returned by the respective services and how they were integrated:

- **Wikimapia** and **OpenStreetMaps** – Both resources return landmarks in the queried area as an XML file that specifies GPS coordinates and additional information. The XML files are parsed using a JavaScript implementation of a SAX parser. The coordinates and names of landmarks are then used to add objects within the Google Maps API.
- **Wikipedia** – In order to integrate landmarks from Wikipedia, we make use of a community created tool called *search-a-place*[6], which returns landmarks from Wikipedia in a given radius of a GPS coordinate. The results are returned in an HTML table that is converted to an XML file similar to the output of Wikimapia. Both the query and the conversion are implemented in a *Yahoo! Pipe*[7] that can be accessed in JavaScript via its URL.
- **Google AJAX Search** – The results returned by the Google AJAX Search API are JavaScript objects that can be directly inserted in the visualisation using the Google Maps API.

3 Using Corpora for Generation

A data-driven generation approach achieves a number of advantages over traditional approaches for our scenario. First of all, corpus data can be used to learn directly how certain events are typically expressed in natural language, thus avoiding the need of manually specifying linguistic realisations. Secondly, variations of discourse structures found in naturally given directions can be learned and reproduced to avoid monotonous descriptions in the generation part. Last but not least, a corpus with good coverage can help us determine the correct selection restrictions on verbs and nouns occurring in directions. The price to pay for these advantages is the cost of annotation; however we believe that this is a reasonable trade-off, in view of the fact that a small annotated corpus and reasonable generalizations in data modelling will likely yield enough information for the intended navigation applications.

3.1 Data Collection

We currently use the data set from (Marciniak & Strube, 2005) to learn linguistic expressions for our generation approach. The data is annotated on the following levels:

- Token and POS level
- Grammatical level (including annotations of main verbs, arguments and connectives)
- Frame-semantics level (including semantic roles and frame annotations in the sense of (Fillmore, 1977))
- Temporal level (including temporal relations between discourse units)

3.2 Our Generation Approach

At the time of writing, our system only makes use of the first three annotation levels. The lexical selection is inspired by the work of Ratnaparkhi (2000) with the overall process designed as follows: given a certain situation on a route, our generation component receives the respective frame name and a list of semantic role filling landmarks as input (cf. Section 4). The generation component then determines a list of potential lexical items to express this frame using the relative frequencies of verbs annotated as evoking the particular frame with the respective set of semantic roles (examples in Table 1).

SELF_MOTION	
PATH	17% *walk*, 13% *follow*, 10% *cross*, 7% *continue*, 6% *take*, …
GOAL	18% *get*, 18% *enter*, 9% *continue*, 7% *head*, 5% *reach*, …
SOURCE	14% *leave*, 14% *start*, …
DIRECTION	25% *continue*, 13% *make*, 13% *walk*, 6% *go*, 3% *take*, …
DISTANCE	15% *continue*, 8% *go*, ...
PATH + GOAL	29% *continue*, 14% *take*, ...
DISTANCE + GOAL	100% *walk*
DIRECTION + PATH	23% *continue*, 23% *walk*, 8% *take*, 6% *turn*, 6% *face*, …

Table 1: Probabilities of lexical items for the frame SELF_MOTION and different frame elements

[6] http://toolserver.org/~kolossos/wp-world/umkreis.php

[7] http://pipes.yahoo.com/pipes/pipe.info?_id=BBI0x8G73RGbWzKnBR50VA

For frame-evoking elements and each associated semantic role-filler in the situation, the grammatical knowledge learned from the annotation level determines how these parts can be put together in order to generate a full sentence (cf. Table 2).

SELF_MOTION	
walk + [*building* $_{\text{PATH}}$]	*walk* → *walk* + PP PP → along + NP NP → the + *building*
get + [*building* $_{\text{GOAL}}$]	*get* → *get* + to + NP NP → the + *building*
take + [*left* $_{\text{DIRECTION}}$]	*take* → *take* + NP NP → a + *left*

Table 2: Examples of phrase structures for the frame SELF_MOTION and different semantic role fillers

4 Combining Text and Visualisation

As mentioned in the previous section, our model is able to compute single instructions at crucial points of a route. At the time of writing the actual integration of this component consists of a set of hardcoded rules that map route segments to frames, and landmarks within the segment to role fillers of the considered frame. The rules are specified as follows:

- A turning point given by the Google Maps API is mapped to the SELF_MOTION frame with the actual direction as the semantic role *direction*. If there is a landmark adjacent to the turning point, it is added to the frame as the role filler of the role *source*.
- If a landmark is adjacent or within the starting point of the route, it will be mapped to the SELF_MOTION frame with the landmark filling the semantic role *source*.
- If a landmark is adjacent or within the goal of a route, it will be mapped to the SELF_MOTION frame with the landmark filling the semantic role *goal*.
- If a landmark is adjacent to a route or a route segment is within a landmark, the respective segment will be mapped to the SELF_MOTION frame with the landmark filling the semantic role *path*.

5 Conclusions and Outlook

We have presented the technical details of an early research prototype that uses NLG methods to generate walking directions for routes computed by an online route planner. We outlined the advantages of a data-driven generation approach over traditional rule-based approaches and implemented a first-version application, which can be used as an initial prototype extensible for further research and development.

Our next goal in developing this system is to enhance the generation component with an integrated model based on machine learning techniques that will also account for discourse level phenomena typically found in natural language directions. We further intend to replace the current hard-coded set of mapping rules with an automatically induced mapping that aligns physical routes and landmarks with the semantic representations. The application is planned to be used in web experiments to acquire further data for alignment and to study specific effects in the generation of walking instructions in a multimodal setting.

The prototype system described above will be made publicly available at the time of publication.

Acknowledgements

This work is supported by the DFG-financed innovation fund FRONTIER as part of the Excellence Initiative at Heidelberg University (ZUK 49/1).

References

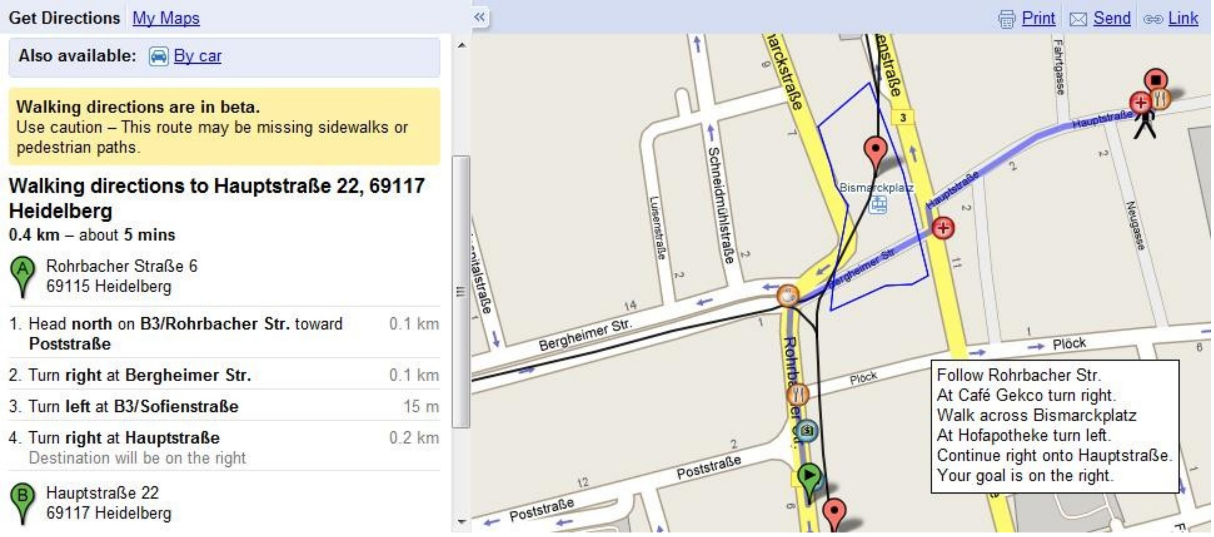
Dale, R., Geldof, S., & Prost, J.-P. (2002). Generating more natural route descriptions. *Proceedings of the 2002 Australasian Natural Language Processing Workshop.* Canberra, Australia.

Fillmore, C. (1977). The need for a frame semantics in linguistics. *Methods in Linguistics , 12*, 2-29.

Marciniak, T., & Strube, M. (2005). Using an annotated corpus as a knowledge source for language generation. *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, (pp. 19-24). Birmingham, UK.

Ratnaparkhi, A. (2000). Trainable Methods for Surface Natural Language Generation. *Proceedings of the 6th Applied Natural Language Processing Conference.* Seattle, WA, USA.

Tom, A., & Denis, M. (2003). Referring to landmark or street information in route directions: What difference does it make? In W. Kuhn, M. Worboys, & S. Timpf (Eds.), *Spatial Information Theory* (pp. 384-397). Berlin: Springer.



Get Directions My Maps

Also available: By car

Walking directions are in beta.
Use caution – This route may be missing sidewalks or pedestrian paths.

Walking directions to Hauptstraße 22, 69117 Heidelberg
0.4 km – about **5 mins**

A Rohrbacher Straße 6
69115 Heidelberg

1. Head **north** on **B3/Rohrbacher Str.** toward **Poststraße** 0.1 km
2. Turn **right** at **Bergheimer Str.** 0.1 km
3. Turn **left** at **B3/Sofienstraße** 15 m
4. Turn **right** at **Hauptstraße** 0.2 km
Destination will be on the right

B Hauptstraße 22
69117 Heidelberg

Print Send Link

Follow Rohrbacher Str.
At Café Gekco turn right.
Walk across Bismarckplatz
At Hofapotheke turn left.
Continue right onto Hauptstraße.
Your goal is on the right.

Figure 1: Visualised route from *Rohrbacher Straße 6* to *Hauptstrasse 22, Heidelberg*. Left: GoogleMaps directions; Right: GoogleMaps visualisation enriched with landmarks and directions generated by our system (The directions were manually inserted here as they are actually presented step-by-step following the route)

Script Outline

Our demonstration is outlined as follows: At first we will have a look at the textual outputs of standard route planners and discuss at which points the respective instructions could be improved in order to be better understandable or easier to follow. We will then give an overview of different types of landmarks and argue how their integration into route directions is a valuable step towards better and more natural instructions.

Following the motivation of our work, we will present different online resources that provide landmarks of various sorts. We will look at the information provided by these resources, examine the respective input and output formats, and state how the formats are integrated into a common data representation in order to access the information within the presented application.

Next, we will give a brief overview of the corpus in use and point out which kinds of annotations were available to train the statistical generation component. We will discuss which other annotation levels would be useful in this scenario and which disadvantages we see in the current corpus. Subsequently we outline our plans to acquire further data by collecting directions for routes computed via Google Maps, which would allow an easier alignment between the instructions and routes.

Finally, we conclude the demonstration with a presentation of our system in action. During the presentation, the audience will be given the possibility to ask questions and propose routes for which we show our system's computation and output (cf. Figure 1).

System Requirements

The system is currently developed as a web-based application that can be viewed with any JavaScript supporting browser. A mid-end CPU is required to view the dynamic route presentation given by the application. Depending on the presentation mode, we can bring our own laptop so that the only requirements to the local organisers would be a stable internet connection (access to the resources mentioned in the system description is required) and presentation hardware (projector or sufficiently large display).

Combining POMDPs trained with User Simulations and Rule-based Dialogue Management in a Spoken Dialogue System

Sebastian Varges, Silvia Quarteroni, Giuseppe Riccardi, Alexei V. Ivanov, Pierluigi Roberti
Department of Information Engineering and Computer Science
University of Trento
38050 Povo di Trento, Italy
{varges|silviaq|riccardi|ivanov|roberti}@disi.unitn.it

Abstract

Over several years, we have developed an approach to spoken dialogue systems that includes rule-based and trainable dialogue managers, spoken language understanding and generation modules, and a comprehensive dialogue system architecture. We present a Reinforcement Learning-based dialogue system that goes beyond standard rule-based models and computes on-line decisions of the best dialogue moves. The key concept of this work is that we bridge the gap between manually written dialog models (e.g. rule-based) and adaptive computational models such as Partially Observable Markov Decision Processes (POMDP) based dialogue managers.

1 Reinforcement Learning-based Dialogue Management

In recent years, Machine Learning techniques, in particular Reinforcement Learning (RL), have been applied to the task of dialogue management (DM) (Levin et al., 2000; Williams and Young, 2006). A major motivation is to improve robustness in the face of uncertainty, for example due to speech recognition errors. A further motivation is to improve adaptivity w.r.t. different user behaviour and application/recognition environments. The Reinforcement Learning framework is attractive because it offers a statistical model representing the dynamics of the interaction between system and user. This is in contrast to the supervised learning approach of learning system behaviour based on a fixed corpus (Higashinaka et al., 2003). To explore the range of dialogue management strategies, a simulation environment is required that includes a simulated user (Schatzmann et al., 2006) if one wants to avoid the prohibitive cost of using human subjects.

We demonstrate the various parameters that influence the learnt dialogue management policy by using pre-trained policies (section 4). The application domain is a tourist information system for accommodation and events in the local area. The domain of the trained DMs is identical to that of a rule-based DM that was used by human users (section 2), allowing us to compare the two directly. The state of the POMDP keeps track of the SLU hypotheses in the form of domain concepts (10 in the application domain, e.g. main activity, star rating of hotels, dates etc.) and their values. These values may be abstracted into ‘known/unknown,’ for example, increasing the likelihood that the system re-visits a dialogue state which it can exploit. Representing the verification status of the concepts in the state, influences – in combination with the user model (section 1.2) and N best hypotheses – if the system learns to use clarification questions.

1.1 The exploration/exploitation trade-off in reinforcement learning

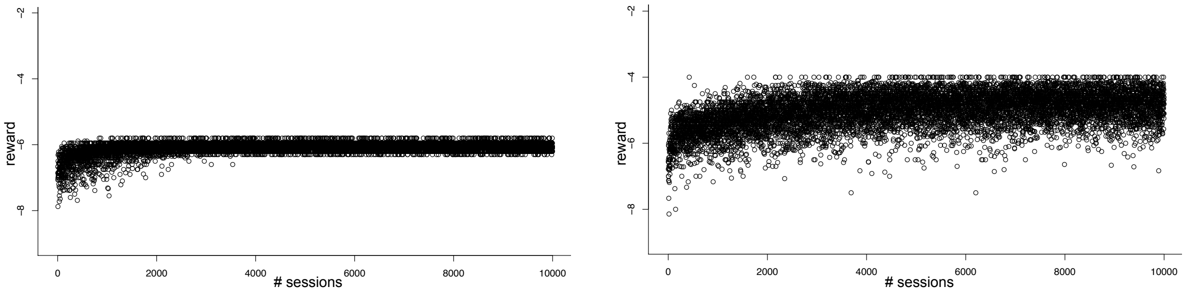
The RL-DM maintains a policy, an internal data structure that keeps track of the values (accumulated rewards) of past state-action pairs. The goal of the learner is to optimize the long-term reward by maximizing the ‘Q-Value’ $Q^{\pi}(s_t, a)$ of a policy π for taking action a at time t. The expected cumulative value V of a state s is defined recursively as $V^{\pi}(s_t) =$

$$\sum_{a} \pi(s_t, a) \sum_{s_{t+1}} P^{a}_{s_t,s_{t+1}} [R^{a}_{s_t,s_{t+1}} + \gamma V^{\pi}(s_{t+1})].$$

Since an analytic solution to finding an optimal value function is not possible for realistic dialogue scenarios, $V(s)$ is estimated by dialogue simulations.

To optimize Q and populate the policy with expected values, the learner needs to explore untried actions (system moves) to gain more experiences, and combine this with exploitation of the

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 41–44,
Suntec, Singapore, 3 August 2009. ©2009 ACL and AFNLP



(a) 0% exploration, 100% exploitation: learner does not find optimal dialogue strategy

(b) 20% exploration, 80% exploitation: noticeable increase in reward, hitting upper bound

Figure 1: Exploration/exploitation trade-off

already known successful actions to also ensure high reward. In principle there is no distinction between training and testing. Learning in the RL-based dialogue manager is strongly dependent on the chosen exploration/exploitation trade-off. This is determined by the action selection policy, which for each system turn decides probabilistically (ϵ-greedy, softmax) if to exploit the currently known best action of the policy for the believed dialogue state, or to explore an untried action. Figure 1(a) shows, for a subdomain of the application domain, how the reward (expressed as minimizing costs) reaches an upper bound early during 10,000 simulated dialogue sessions (each dot represents the average of 10 rewards at a particular session number). Note that if the policy provides no matching state, the system can only explore, and thus a certain amount of exploration always takes place. In contrast, with exploration the system is able to find lower cost solutions (figure 1(b)).

1.2 User Simulation

In order to conduct thousands of simulated dialogues, the DM needs to deal with heterogeneous but plausible user input. For this purpose, we have designed a User Simulator (US) which bootstraps likely user behaviors starting from a small corpus of 74 in-domain dialogs, acquired using the rule-based version of the SDS (section 2). The task of the US is to simulate the output of the SLU module to the DM, hence providing it with a ranked list of SLU hypotheses.

A list of possible user goals is stored in a database table (section 3) using a frame/slot representation. For each simulated dialogue, one or more user goals are randomly selected. The User Simulator's task is to mimic a user wanting to perform such task(s). At each turn, the US mines the previous system dialog act to obtain the concepts required by the DM and obtains the corresponding values (if any) from the current user goal.

The output of the user model proper is passed to an error model that simulates the "noisy channel" recognition errors based on statistics from the dialogue corpus. These concern concept values as well as other dialogue phenomena such as noInput, noMatch and hangUp. If the latter phenomena occur, they are propagated to the DM directly; otherwise, the following US step is to attach plausible confidences to concept-value pairs, also based on the dialogue corpus. Finally, concept-value pairs are combined in an SLU hypothesis and, as in the regular SLU module, a cumulative utterance-level confidence is computed, determining the rank of each of the n hypotheses. The probability of a given concept-value observation at time $t+1$ given the system act at time t, named $a_{s,t}$, and the session user goal g_u, $P(o_{t+1}|a_{s,t}, g_u)$, is obtained by combining the error model and the user model:

$$P(o_{t+1}|a_{u,t+1}) \cdot P(a_{u,t+1}|a_{s,t}, g_u)$$

where $a_{u,t+1}$ is the true user action.

2 Rule-based Dialogue Management

A rule-based dialogue manager was developed as a meaningful comparison to the trained DM, to obtain training data from human-system interaction for the user simulator, and to understand the properties of the domain (Varges et al., 2008). Rule-based dialog management works in two stages: retrieving and preprocessing facts (tuples) taken from a dialogue state database (section 3), and inferencing over those facts to generate a system response. We distinguish between the 'context model' of the first phase – essentially allowing more recent values for a concept to override less recent ones – and the 'dialog move engine' (DME) of the second phase. In the second stage, acceptor rules match SLU results to dialogue context, for example perceived user concepts to open questions. This may result in the decision to verify the application parameter in question, and the action is verbalized by language generation rules. If the parameter is accepted, application dependent task rules determine the next parameter to be acquired, resulting in the generation of an appropriate request.

3 Data-centric System Architecture

All data is continuously stored in a database which web-service based processing modules (such as SLU, DM and language generation) access. This architecture also allows us to access the database for immediate visualization. The system presents an example of a "thick" inter-module information pipeline architecture. Individual components exchange data by means of sets of hypotheses complemented by the detailed conversational context. The database concentrates heterogeneous types of information at various levels of description in a uniform way. This facilitates dialog evaluation, data mining and online learning because data is available for querying as soon as it has been stored. There is no need for separate logging mechanisms. Multiple systems/applications are available on the same infrastructure due to a clean separation of its processing modules (SLU, DM, NLG etc.) from data storage (DBMS), and monitoring/analysis/visualization and annotation tools.

4 Visualization Tool

We developed a live web-based dialogue visualization tool that displays ongoing and past dialogue utterances, semantic interpretation confidences and distributions of confidences for incoming user acts, the dialogue manager state, and policy-based decisions and updating. An example of the visualization tool is given in figures 3 (dialogue logs) and 4 (annotation view). We are currently extending the visualization tool to display the POMDP-related information that is already present in the dialogue database.

The visualization tool shows how our dedicated SLU module produces a number of candidate semantic parses using the semantics of a domain ontology and the output of ASR.

The visualization of the internal representation of the POMDP-DM includes the N best dialogue states after each user utterance and the reranking of the action set. At the end of each dialogue session, the reward and the policy updates are shown, i.e. new or updated state entries and action values. Another plot relates the current dialogue's reward to the reward of previous dialogues (as in plots 1(b) and 1(a)).

Users are able to talk with several systems (via SIP phone connection to the dialogue system server) and see their dialogues in the visualization tool. They are able to compare the rule-based system, a randomly exploring learner that has not been trained yet, and several systems that use various pre-trained policies. These policies are obtained by dialogue simulations with user models based on data obtained from human-machine dialogues with the original rule-based dialogue manager. The web tool is available at `http://cicerone.dit.unitn.it/DialogStatistics/`.

Acknowledgments

This work was partially supported by the European Commission Marie Curie Excellence Grant for the ADAMACH project (contract No. 022593) and by LUNA STREP project (contract No. 33549).

References

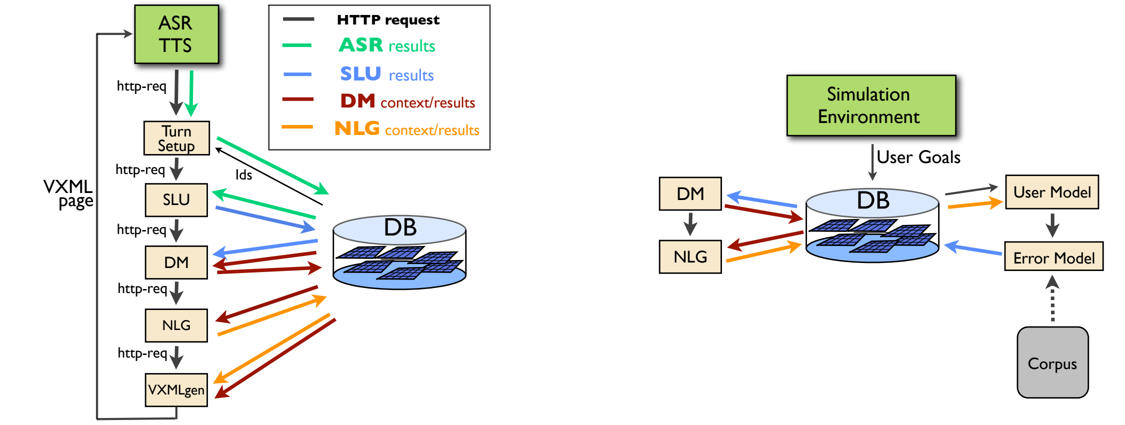
R. Higashinaka, M. Nakano, and K. Aikawa. 2003. Corpus-based discourse understanding in spoken dialogue systems. In *ACL-03*, Sapporo, Japan.

E. Levin, R. Pieraccini, and W. Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1).

J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. 2006. A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. *Knowledge Engineering Review*, 21(2):97–126.

S. Varges, G. Riccardi, and S. Quarteroni. 2008. Persistent Information State in a Data-Centric Architecture. In *SIGDIAL-08*, Columbus, Ohio.

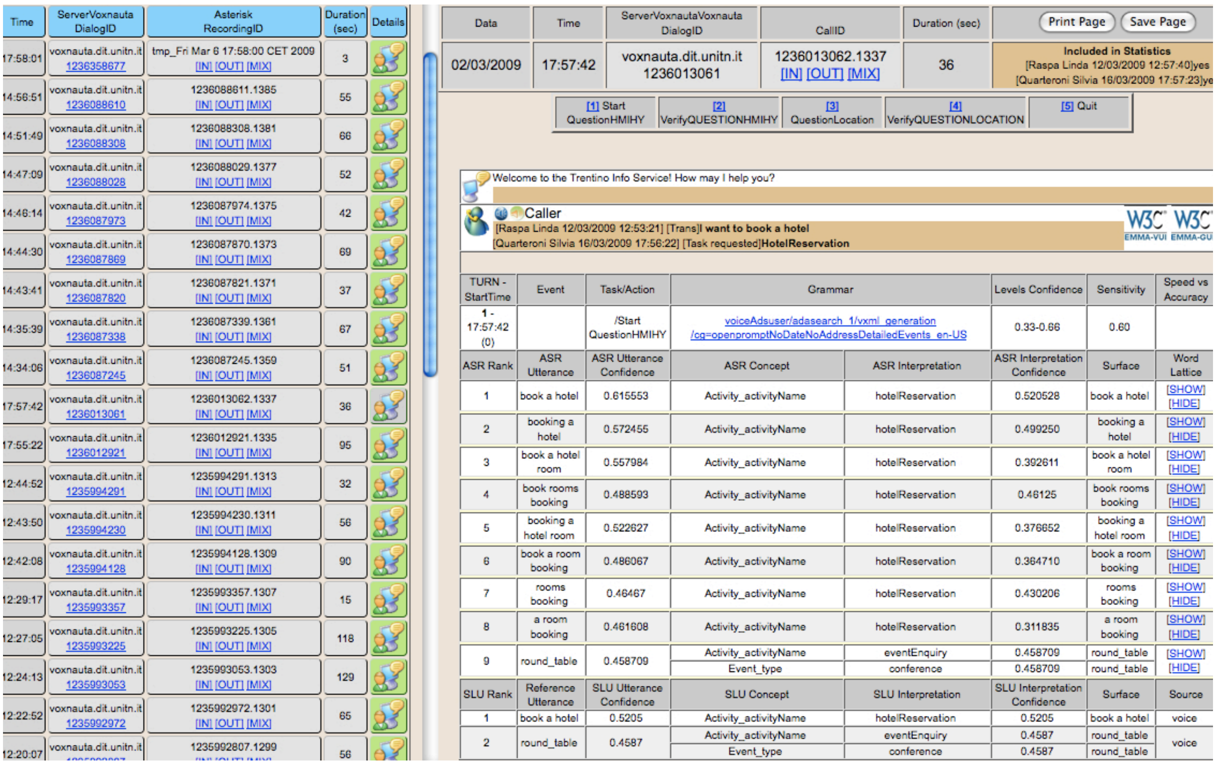
J. D. Williams and S. Young. 2006. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language*, 21(2):393–422.



(a) Turn-level information flow in the data-centric SDS architecture

(b) User simulator interface with the dialogue manager

Figure 2: Architecture for interacting with human user (left) and simulated user (right)



Time	ServerVoxnauta DialogID	Asterisk RecordingID	Duration (sec)	Details
17:58:01	voxnauta.dit.unitn.it 1236358677	tmp_Fri Mar 6 17:58:00 CET 2009 [IN] [OUT] [MIX]	3	
14:56:51	voxnauta.dit.unitn.it 1236088610	1236088611.1385 [IN] [OUT] [MIX]	55	
14:51:49	voxnauta.dit.unitn.it 1236088308	1236088308.1381 [IN] [OUT] [MIX]	66	
14:47:09	voxnauta.dit.unitn.it 1236088028	1236088029.1377 [IN] [OUT] [MIX]	52	
14:46:14	voxnauta.dit.unitn.it 1236087973	1236087974.1375 [IN] [OUT] [MIX]	42	
14:44:30	voxnauta.dit.unitn.it 1236087869	1236087870.1373 [IN] [OUT] [MIX]	69	
14:43:41	voxnauta.dit.unitn.it 1236087820	1236087821.1371 [IN] [OUT] [MIX]	37	
14:35:39	voxnauta.dit.unitn.it 1236087338	1236087339.1361 [IN] [OUT] [MIX]	67	
14:34:06	voxnauta.dit.unitn.it 1236087245	1236087245.1359 [IN] [OUT] [MIX]	51	
17:57:42	voxnauta.dit.unitn.it 1236013061	1236013062.1337 [IN] [OUT] [MIX]	36	
17:55:22	voxnauta.dit.unitn.it 1236012921	1236012921.1335 [IN] [OUT] [MIX]	95	
12:44:52	voxnauta.dit.unitn.it 1235994291	1235994291.1313 [IN] [OUT] [MIX]	32	
12:43:50	voxnauta.dit.unitn.it 1235994230	1235994230.1311 [IN] [OUT] [MIX]	56	
12:42:08	voxnauta.dit.unitn.it 1235994128	1235994128.1309 [IN] [OUT] [MIX]	90	
12:29:17	voxnauta.dit.unitn.it 1235993357	1235993357.1307 [IN] [OUT] [MIX]	15	
12:27:05	voxnauta.dit.unitn.it 1235993225	1235993225.1305 [IN] [OUT] [MIX]	118	
12:24:13	voxnauta.dit.unitn.it 1235993053	1235993053.1303 [IN] [OUT] [MIX]	129	
12:22:52	voxnauta.dit.unitn.it 1235992972	1235992972.1301 [IN] [OUT] [MIX]	65	
12:20:07	voxnauta.dit.unitn.it	1235992807.1299	56	

Data	Time	ServerVoxnautaVoxnauta DialogID	CallID	Duration (sec)	Print Page Save Page
02/03/2009	17:57:42	voxnauta.dit.unitn.it 1236013061	1236013062.1337 [IN] [OUT] [MIX]	36	**Included in Statistics** [Raspa Linda 12/03/2009 12:57:40]yes [Quarteroni Silvia 16/03/2009 17:57:23]ye

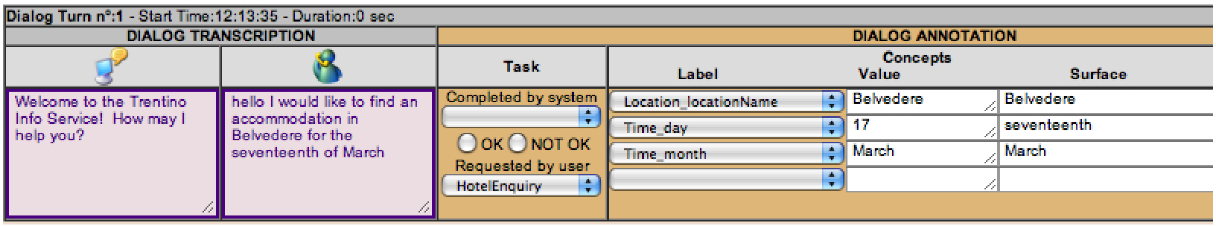
[1] Start QuestionHMIHY	[2] VerifyQUESTIONHMIHY	[3] QuestionLocation	[4] VerifyQUESTIONLOCATION	[5] Quit

Welcome to the Trentino Info Service! How may I help you?

Caller
[Raspa Linda 12/03/2009 12:53:21] [Trans]**I want to book a hotel**
[Quarteroni Silvia 16/03/2009 17:56:22] [Task requested]**HotelReservation**

TURN - StartTime	Event	Task/Action	Grammar		Levels Confidence	Sensitivity	Speed vs Accuracy
1 - 17:57:42 (0)		/Start QuestionHMIHY	voiceAdsuser/adasearch_1/vxml_generation /cg=openpromptNoDateNoAddressDetailedEvents_en-US		0.33-0.66	0.60	
ASR Rank	ASR Utterance	ASR Utterance Confidence	ASR Concept	ASR Interpretation	ASR Interpretation Confidence	Surface	Word Lattice
1	book a hotel	0.615553	Activity_activityName	hotelReservation	0.520528	book a hotel	[SHOW] [HIDE]
2	booking a hotel	0.572455	Activity_activityName	hotelReservation	0.499250	booking a hotel	[SHOW] [HIDE]
3	book a hotel room	0.557984	Activity_activityName	hotelReservation	0.392611	book a hotel room	[SHOW] [HIDE]
4	book rooms booking	0.488593	Activity_activityName	hotelReservation	0.46125	book rooms booking	[SHOW] [HIDE]
5	booking a hotel room	0.522627	Activity_activityName	hotelReservation	0.376652	booking a hotel room	[SHOW] [HIDE]
6	book a room booking	0.486067	Activity_activityName	hotelReservation	0.364710	book a room booking	[SHOW] [HIDE]
7	rooms booking	0.46467	Activity_activityName	hotelReservation	0.430206	rooms booking	[SHOW] [HIDE]
8	a room booking	0.461608	Activity_activityName	hotelReservation	0.311835	a room booking	[SHOW] [HIDE]
9	round_table	0.458709	Activity_activityName Event_type	eventEnquiry conference	0.458709 0.458709	round_table round_table	[SHOW] [HIDE]
SLU Rank	Reference Utterance	SLU Utterance Confidence	SLU Concept	SLU Interpretation	SLU Interpretation Confidence	Surface	Source
1	book a hotel	0.5205	Activity_activityName	hotelReservation	0.5205	book a hotel	voice
2	round_table	0.4587	Activity_activityName Event_type	eventEnquiry conference	0.4587 0.4587	round_table round_table	voice

Figure 3: Left pane: overview of all dialogues. Right pane: visualization of a system opening prompt followed by the user's activity request. All *distinct* SLU hypotheses (concept-value combinations) deriving from ASR are ranked based on concept-level confidence (2 in this turn).



Dialog Turn n°:1 - Start Time:12:13:35 - Duration:0 sec

DIALOG TRANSCRIPTION		DIALOG ANNOTATION			
		Task	Concepts Label	Value	Surface
Welcome to the Trentino Info Service! How may I help you?	hello I would like to find an accommodation in Belvedere for the seventeenth of March	Completed by system ○ OK ○ NOT OK Requested by user HotelEnquiry	Location_locationName Time_day Time_month	Belvedere 17 March	Belvedere seventeenth March

Figure 4: Turn annotation of task success based on previously filled dialog transcriptions (left box).

Author Index