Proceedings of the International Multiconference on
Computer Science and Information Technology, pp. 339 – 344

ISBN 978-83-60810-14-9
ISSN 1896-7094

Smarty–Extendable Framwork for Bilingual and Multilingual Comprehension Assistants

Todor Arnaudov
Plovdiv University "Paisii Hilendarski", 24, "Tsar Assen" Str., Plovdiv 4000, Bulgaria
Email: tosh.bg@gmail.com

Ruslan Mitkov
University of Wolverhampton, Wolverhampton WV1 1SB, UK
Email: r.mitkov@wlv.ac.uk

***Abstract*—This paper discusses a framework for development of bilingual and multilingual comprehension assistants and presents a prototype implementation of an English-Bulgarian comprehension assistant. The framework is based on the application of advanced graphical user interface techniques, WordNet and compatible lexical databases as well as a series of NLP preprocessing tasks, including POS-tagging, lemmatisation, multiword expressions recognition and word sense disambiguation. The aim of this framework is to speed up the process of dictionary look-up, to offer enhanced look-up functionalities and to perform a context-sensitive narrowing-down of the set of translation alternatives proposed to the user.**

I. Introduction

AT PRESENT, even regular Internet users often access on-line resources in English which require lexical knowledge beyond their current level.

While Machine Translation has been expected to make this problem history, the state-of-the-art is still far from achieving this dream. Full-text machine translation is yet unreliable, and typical users are assisted in translation and language learning with only a variety of word-translation 'electronic dictionaries', operating either on-line on the Internet, or as off-line computer software. Generally, such dictionaries offer very simple look-up options and are based on the following functionalities:

1. The u ser types or copy-pastes a word in the input box, or clicks on a word from an alphabetical list of words.

2.The dictionary displays an entry if it contains one whose head word exactly matches the word which is entered. In most cases, the entry is from a scanned version of a paper dictionary. It is not difficult to see that the way a user consults an electronic dictionary is not very different from the way s/he queries paper dictionaries. As with paper dictionaries, the user is presented a list of possible meanings for every word under consideration. In many cases this could cause confusion or misunderstanding. Recent years have seen the development of new lexicographic/language learners' tools referred to as *comprehension assistants* which seek to enhance the look-up functionality and in particular to narrow down the list of alternative translations through applying basic NLP techniques.

II. Previous Work

A. Xerox

The first comprehension assistant reported, *Locolex* (Feldweg and Breidt 1996), was developed by Xerox for French-English and English-German comprehension assistance. Locolex inspired applications developed later including Smarty, which is being discussed in this paper. Locolex, unlike conventional electronic dictionaries, offers the functionality for the user to click on words occurring in any machine-readable text, as opposed to copy-and-pasting separate words. Once the user clicks on a specific word, Locolex performs POS tagging which attempts to identify the correct part-of-speech tag, thus decreasing the number of possible translations. Multiword expressions recognition, based on regular expressions, is also applied, which could help identify the correct translation in particular cases. Locolex also keeps record of user sessions, allowing quick recall of previously checked words.

A recent version of Locolex incorporates word sense disambiguation which contributes to the narrowing down of the set of possible meanings even further.

B. *Morphologic*

The comprehension assistants developed by Morphologic introduce several additional features.

In particular, one of their products, *MobiMouse* , correctly identifies multiword expressions even if the selected word is not the head of the expression and also offers comprehension assistance in any application running in the operating system environment. The user can click anywhere; the comprehension assistant is running in the background and flashes a translation in the corner of the screen.

C. *SmartDict*

SmartDict (Kolev, 2005) is an English-Bulgarian dictionary which is somewhere between comprehension assistants and advanced conventional dictionaries. It performs a number of NLP preprocessings, including tokenisation, sentence-splitting, normalisation and multiword expression recognition, but it does not perform reduction of the possible translations by POS-tagging or word-sense disambiguation.

978-83-60810-14-9/08/$25.00 © 2008 IEEE

III. Smarty – framework for bilingual and multilingual comprehension assistants

Inspired by Locolex, we developed Smarty - a framework for comprehension assistants for English-Bulgarian. While Smarty and Locolex share certain similarities, our comprehension assistant has the following distinctive features.

A. New Features

While Smarty and Locolex share certain similarities, our comprehension assistant has the following distinctive features.

1) Hybrid System

Smarty represents a hybrid system. The interface is more comprehensive and elaborate than the interface of Locolex or MobiMouse in that it allows users to virtually work with two dictionaries – both an enhanced conventional dictionary and a comprehension assistant (see Fig. 1). In enhanced conventional dictionary mode users can browse freely all dictionary entries and familiarise themselves with the meanings of a specific word. This mode offers additional options such as suffix search, rhyme search, synonymy search etc. which are not present in conventional dictionaries.

2) New Lexicographical Resource

WordNet (Miller, 1995) is the lexicographical resource for this comprehension assistant. WordNet adds glosses which can be browsed by the user. Additionally, it makes it possible for word sense disambiguation to be performed.

3) Extendability

The alignment between the lexical database of WordNet and corresponding lexical databases for other languages allows bilingual word sense disambiguation to be performed. The incorporation of the existing databases of EuroWordNet (Vossen, 1998) and BalkaNet (Oflazer et. al, 2001), make it perfectly possible for comprehension assistants covering more languages to be developed using the same framework and the same core system. Smarty could be easily extended to be English to Greek, Turkish, Czech, Romanian, Serbian, Italian, Spanish, German, French, Dutch and Estonian comprehension assistants, if their already made lexical databases are available.

B. Graphical User Interface

The aspects of the graphical user interface in the framework, which are different from the framework of conventional dictionaries, are:

1) Free Text Input

There is a free text input box, where the full text is pasted or typed. Users can point to the words in their context, instead of copy-pasting (Fig. 2).

2) Tooltip

Suggested translated meanings can appear in a tooltip, near the mouse pointer. This is less distracting for users than the translation appearing in a side window.

3) Additional Information

Additional information which assists comprehension is available - glosses, examples of usage etc. and can also be presented to the user in tooltip or in side windows, on demand.

Translation in tooltip proves to be much more convenient for users than translation in separate windows in the same applications or in the worse case - in another application.

POS-tagging allows Smarty to fit the most relevant translations in a tooltip, which could be scanned in few seconds by the user, without touching a scroller and without moving his or her sight away from the context. This also allows immediate continuation of the reading without distraction.

When using a conventional bilingual dictionary, if the queried word has ambiguous part-of-speech and a long entry with sections for each one, the user faces two problems: s/he is forced to figure out the correct part of speech alone; and if the user knows the part-of-speech s/he is forced to scroll and scan with the bare eye where the section for the correct part-of-speech begins. Also, the appearance of the translation in window of another application causes two other time penalties: first, the user is distracted from the reading flow and has to spend time switching attention from the text to the dictionary and back; and second, after the query is done, the user has to find the exact place in the text window where he or she has stopped reading.

C. Linguistic Databases

The framework makes use of at least three linguistics databases: a conventional dictionary database and at least two lexical databases used to provide glosses and word sense disambiguation.

1) Conventional Dictionary Database

The conventional dictionary database allows the system to work in conventional dictionary mode. It could be a scan of a paper dictionary, which in this case is to be parsed and transformed in suitable format for processing. This database is used to build indices for *predictive typing* (known also as *autocompletion*), suffix-search, rhyme search etc.

An English-Bulgarian dictionary database (a scan of a paper dictionary with about 51000 entries) was used in Smarty because it was freely available and suited for the purpose of this prototype.

Some of the entries include examples of usage, multiword expressions and phrasal verbs, which are parsed and used as resources for multiword expression recognition.

2) WordNet

WordNet is a large lexical database, consisting of synonym sets of words – "synsets" –s tructured by part-of-speech and numerous types of semantic relations. The richness of its structural information makes it a highly acceptable resource for various NLP tasks (Mitkov, 2003). In the proposed framework, it provides glosses, which are used as semantic database for word sense disambiguation (WSD).

Semantic relations included in WordNet – hyperonymy, meronymy, synonymy etc. – could be used in future versions to improve the precision of the WSD.

3) EuroWordNet, BalkaNet etc.

EuroWordNet is a multilingual set of semantic databases for European languages, which are aligned to WordNet and to each other. It consists of databases for Dutch, Italian, Spanish, French, German, Czech and Estonian.

BalkaNet is a similar set, including Bulgarian, Greek, Romanian, Serbian and Turkish lexical databases.

The links between the lexical databases enable direct translation of specific senses. It also allows multilingual translation within a single framework.

In the implementation discussed in this paper, a small version of Bulgarian BalkaNet is used, consisting of about 15000 synsets. However, the system could easily be extended with other databases from the BalkaNet or EuroWordNet frameworks, thus making it possible for "Smarty" to operate as an English-Greek, English-Romanian, English-Serbian etc. comprehension assistant.

D. Natural Language Processing Stages

1) POS-tagging

Selecting a word in context, instead of copy-pasting or typing in a text box allows POS-tagging to be performed. For languages which exhibit typical ambiguity of lexical categories such as English, this could narrow the set of returned dictionary entries by two or three times.

The POS-tagger in Smarty prototype is SharpNLP – an LGPL .NET library, which was chosen because the system is coded in C#.

2) Lemmatisation and Normalisation

This stage saves the user the trimming of words copied from texts and thus speeds up the look-up. Lemmatisation and normalisation are also used in the multiword expression recognition and word-sense disambiguation stages to allow capturing variations.

3) Multiword expressions recognition

At this stage the context of the selected word is checked for matches with multiword expressions in the conventional dictionary database. The words from the context are lemmatised and then fuzzy-matched to patterns from a multiword expressions database. Different techniques are applied to compute the degree of match: bag of words, POS-matching, regular expressions. The fuzzy matching algorithm applied imply high recall, still delivering only one or few most relevant multiword expressions, which fit in a tooltip or a small text box.

Automatic multiword expressions recognition capabilities of Smarty allow faster look-up of multiword expressions, compared to the operation of conventional dictionaries, which usually lack such functionalities or capture only exact matches. A sorted list of multiword expressions and phrasal verbs, presented in Smarty, also helps users quickly find wanted translations of multiword expressions.

Using conventional dictionaries, finding out that there is a multiword expression in certain contexts may require the user to scroll and scan the whole dictionary entry by sight. It must be pointed out, that in cases of words with short entries Smarty does not have a significant advantage because a visual scan of possible expressions could also be done in few seconds. However, the advantage of multiword expressions recognition is significant when quering entries of common words like "run", "take", "go", "have" etc., which have many tenths of examples of usage.

4) Word-Sense Disambiguation

The ultimate goal of comprehension assistants is to find the most appropriate translation in a given context. Word sense disambiguation contributes to the further narrowing down of the list of possible senses. Figure 3 illustrates how the selected word initially featuring 80 potential meanings has the number of its possible translations reduced to 21 after POS tagging and even further reduced to 1 single possible meaning after correct word sense disambiguation.

In this implementation Smarty uses glosses from WordNet to perform simple WSD in English, related to the method of Lesk (Lesk, 1986) . The framework benefits from the alignment between the lexical databases of WordNet and BalkaNet, which allows word-sense disambiguated sense in English to be mapped directly to precise sense in Bulgarian or other language from BalkaNet or EuroWordNet.

The method for WSD in the prototype of Smarty applies the following algorithm:

1. A word in a text is pointed and then its context is tokenized, normalised and POS-tagged. It is then cleaned from stop-words which are considered to be confusing for the process of WSD.
2. WordNet synonym sets corresponding to the queried word are found in the database and their glosses are extracted.
3. Each gloss is tokenised, part-of-speech tagged, lemmatised/normalised and cleaned from stop-words.
4. The normalised context and the gloss are matched and word-matches are counted.
5. Until there are more glosses, go back to step 3. Otherwise:
6. The gloss with highest number of matches is suggested as the most probable sense. If there are not any matches, the most frequent sense referring to WordNet is suggested. If there are more than one senses with the same number of matches, the most frequent sense is suggested also, again referring to the order of senses in WordNet.
7. The index of the synonym set of the suggested sense is matched to the indices of BalkaNet.
8. If BalkaNet contains the disambiguated sense, then disambiguated translation in Bulgarian is displayed with confidence. Otherwise, other available senses are displayed with a sign of uncertainty.

This algorithm has low precision, due to its simplicity. Disambiguation in English is correct in two cases. The first case is when the context of the queried word includes specific discriminative words from the gloss of the correct sense. The second case is when discriminative words are not present in the context, but the most frequent sense is used, because the most frequent sense is suggested in case of uncertainty.

Precision in WSD to Bulgarian is lower than the precision in English, due to the limited size of the lexical database used–15000 synsets [Windows U1] versus 115000. Also, in most cases BalkaNet includes only one or few most frequent senses for a given word.

Examples of correctly disambiguated senses follow. The words from the glosses which are used to discriminate the sense are underlined.

- *What <u>instrument</u> do you play, Paul?*

- *I play the* ***bass.***

Suggested sense: bass – n. the member with the lowest range of a family of musical instruments.

– You are fired! – said the boss.

Suggested sense: fire – v. terminate the employment of; "The boss fired his secretary today"; "The company terminated 25% of its workers".

IV. User Evaluation

Some aspects of Smarty's performance and features were evaluated in real environment by users and compared with two other electronic dictionaries - *SA Dictionary* and *Babylon*.

SA Dictionary is a popular English-Bulgarian conventional dictionary, based on the standard simple framework. *Babylon* is an advanced multilingual conventional dictionary framework with graphical user interface having certain similarities to the interface of comprehension assistants – the system captures words clicked anywhere on the screen. Babylon can recognise[Windows U1] phrasal verbs and multiword expressions, but only if they are in the exact form as they appear in the dictionary database. Also, the dictionary lacks NLP functionality for reducing the number of possible translations of single words.

Babylon offers machine translation, however it employs third-party on-line services (probably Systran) and this specific functionalities are not relevant for this evaluation.

A. Query time for single words translation

Smarty framework provides three main quick results in tooltips:

1. The most relevant part-of-speech portion of the entry from a conventional dictionary.
2. A suggested multiword expression which matches the context of the pointed word.
3. Suggested word-sense disambiguated sense in Bulgarian.

A bubble with either a translation from these types appears virtually immediately on the test PC with 1.8 GHz Athlon CPU. The query time is between 0.2 sec to 1.2 sec. This is where the worst cases are met when querying words with the longest list of multiword expressions, due to the time needed to match them to the context.

SA Dictionary also provides virtually immediate results for single-word queries, while Babylon is delayed by a few seconds due to access to Internet resources. However, both lack the capability to reduce the entries to the most relevant sections. This often slows down the time for actual translation as the user is forced to scan long entries with the bare eye.

A small test was conducted in order to assess the time saved with Smarty (if any) in a real environment. Several chapters from Dan Brown's *The Da Vinci Code* were selected, in order to represent a common text with similar style and language complexity. Two native-Bulgarian speakers with different English proficiency read three chapters with Smarty, SA Dictionary and Babylon. Users queried unknown or ambiguous words, and searched their meanings in the entry displayed in dictionary's window (SA Dictionary and Babylon) or in the tooltip, provided by Smarty. The number of queries, the total time needed for the look-up in seconds and the average time per query were computed.

Table I .
User 1 - Undergraduate student

Chapter	Dictionary	Words	Queries	Time	**Average**
2	SA	909	31	153	**4.94 s/q**
45	Babylon	1149	23	173	**7.52 s/q**
100	Smarty	1183	24	80	**3.33 s/q**

Table II .
User 2 – PhD student

Chapter	Dictionary	Words	Queries	Time	**Average**
2	SA	909	29	74	**2.55 s/q**
45	Babylon	1149	28	100	**3.57 s/q**
100	Smarty	1183	37	97	**2.62 s/q**

The tables show that the PhD student is much faster than the undergraduate student with all three dictionaries. The results also suggest, that in the experiments carried out by the PhD student,Smarty and SA Dictionary are practically equal in performance, while the undergraduate student translates significantly faster with Smarty. Generally both observations could be explained by the higher English proficiency of the PhD student. The equal speed of operation using Smarty and a conventional] dictionary in one of the tests could be explained by the genre of the texts and by the high English proficiency of the PhD student. Her queries consist of rare words, which are not ambiguous , thus the entries in the conventional dictionary could also be scanned in a moment.

We conjecture that Smarty would perform much faster than conventional dictionaries if tested by language learners with much lower English proficiency. Language learners are expected to query more frequent words, which exhibit higher lexical and part-of-speech ambiguity. This is where Smarty's NLP preprocessing can offer significant advantage over the simple operation of conventional dictionaries, and where Smarty would be most useful.

Acknowledgment

We would like to express our gratitude to Constantin Orasan, Irina Temnikova and Kiril Kolev for their suggestions and taking part in the evaluation experiments.

References

[1] H. Feldweg, E. Breidt, "COMPASS An Intel-ligent Dictionary System for Reading Text in a Foreign Language". In F. Kiefer and G. Kiss, editors, Papers in Computational Lexicography, COMPLEX '96, Budapest, pages 53-62.

[2] D. Kolev, "Computer assistant for translators" - Bachelor Thesis (In Bulgarian). Plovdiv University 2005.

[3] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone." In Proc. of the 1986 SIGDOC Conference, pages 24–6, Ontario, Canada.

[4] G. A. Miller, "WordNet: a lexical database for English", Communication of the ACM, Volume 38, Issue 11 (November 1995)

[5] R. Mitkov, R., "The Oxford Handbook of Computa-tional Linguistics", Oxford University Press, 2003.
[6] K. Oflazer, S. Stamou, D. Christodoulakis, “BALKANET: A Multilingual Semantic Network for the Balkan Languages“ – in the Elsnet Newsletter of the European Network in Human Language Technology, 2001.
[7] G. Prószéky, “Comprehension Assistance Meets Machine Translation”. In: Toma3 Erjavec; Jerneja Gros (eds) Language Technologies, 1–5. Institut Jo3ef Stefan, Ljubljana, Slovenia.
[8] P. Vossen, “EuroWordNet: a multilingual data-base with lexical semantic networks”, Kluwer Aca-demic Publishers Norwell, MA, USA, 1998.

APPENDIX

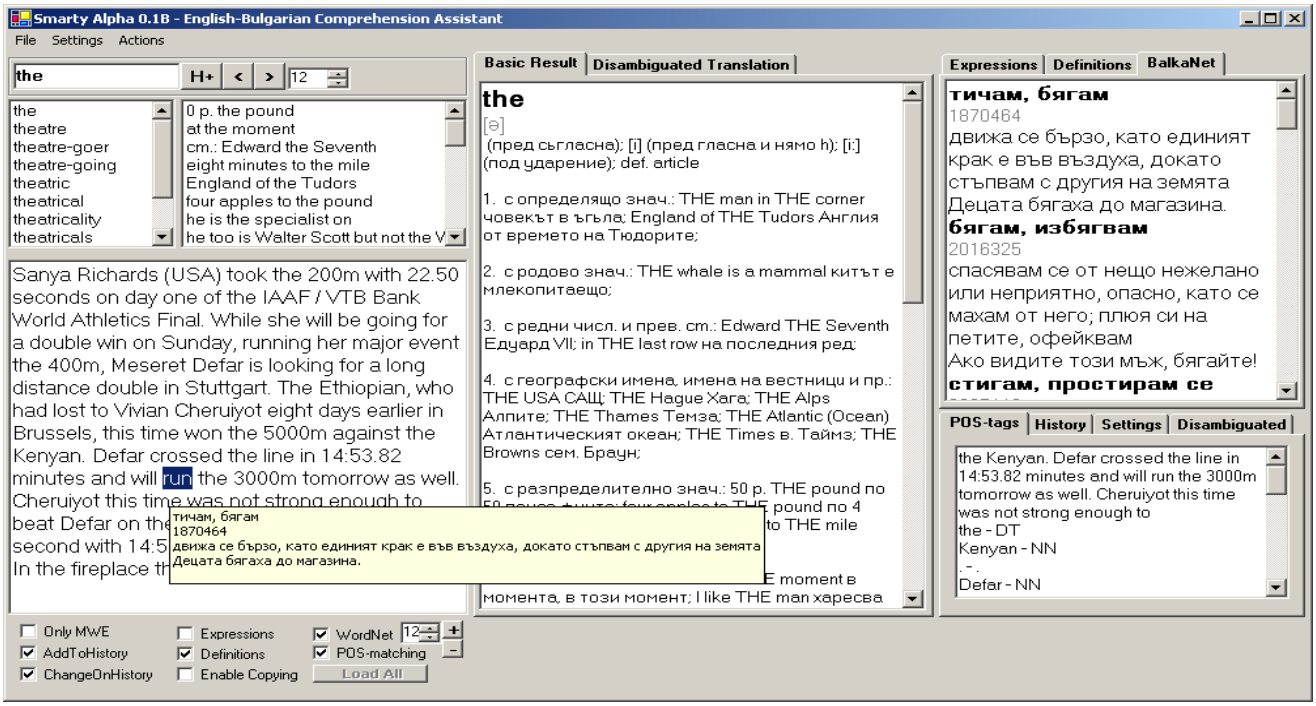

Fig 1 . Smarty

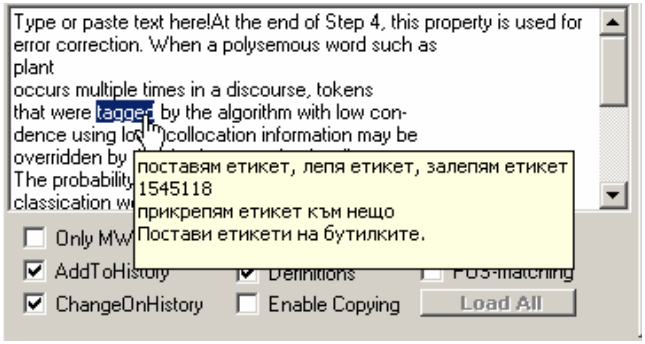

Fig 2 . Bilingual word-sense disambiguation

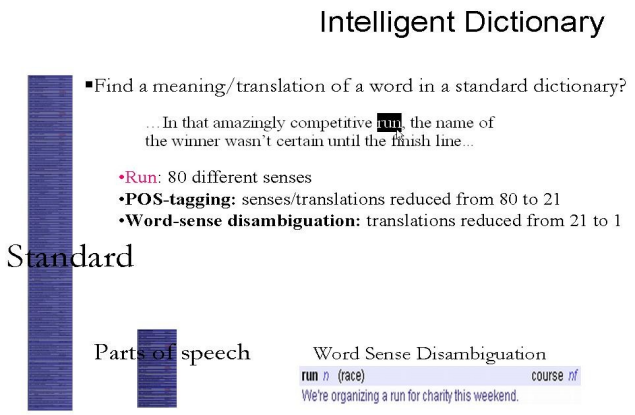

Fig 3 . The process of narrowing the list of possible translations