Smarty – Extendable Framework for Bilingual and Multilingual Comprehension Assistants

Todor Arnaudov*, Ruslan Mitkov**

* Plovdiv University, Email todprog@yahoo.com
** University of Wolverhampton, Email r.mitkov@wlv.ac.uk

Abstract

This paper discusses a framework for development of bilingual and multilingual comprehension assistants. The framework is based on application of advanced graphical user interface techniques, WordNet and compatible lexical databases as well as a series of NLP pre-processing tasks, including POS-tagging, lemmatisation, multiword expressions matching 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 context-sensitive narrowing of the set of translation alternatives, proposed to the user.

1. Introduction

Nowadays users have the choice to query a variety of 'electronic dictionaries', which operate either in on-line or off-line mode. Most such dictionaries offer very simple look-up options and are based on the following functionalities:

1. The user 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 from the dictionary, if there is 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 she/he queries paper dictionaries. Just like with paper dictionaries, the user is presented for every word under consideration with a list of possible meanings which in many cases 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

2. Previous work

2.1. 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 in conventional electronic dictionaries, offers the functionality of the user clicking on words occurring in any machine-readable text, as opposed to of 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. Multi-word 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 already checked words.

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

2.2. Morphologic

The comprehension assistants developed by Morphologic introduce several additional features.

In particular, one of their products, *Mobi-Mouse*, correctly identifies multi-word expressions even if the selected word is not the head of a multi-word expression and 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.

3. Smarty - framework for bilingual and multilingual comprehension assistants

Inspired by Locolex, we developed *Smarty* - a framework for comprehension assistant for English-Bulgarian. 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 virtually to work with two dictionaries – both an enhanced conventional dictionary and a comprehension assistant (see Figure 1). In an enhanced conventional dictionary mode users can browse freely all dictionary entries and familiarise themselves with the meanings of a specific word, if they wish. 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 is the lexicographical resource for this comprehension assistant. WordNet adds glosses, which can be browsed by the user. Besides, it makes it possible for word sense disambiguation to be performed.
3. *Exendability*: The alignment to WordNet lexical databases allows bilingual word sense disambiguation. The incorporation of the existing databases of EuroWordNet and BalkaNet, makes it perfectly possible for comprehension assistants covering more languages, to be developed using the same framework and the same core system.

3.1. Graphical User Interface

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

1. There is a *free text input box*, where full text is pasted or typed. Users can point to the words in their context, instead of doing copy-pasting (Figure 2).
2. Suggested translated meanings can appear in a *tooltip*, near the mouse pointer. This is less distracting for users than if translation appears in a side window.
3. Additional information which assists comprehension is available - glosses, examples of usage, concordances from corpora etc. and can also be presented to the user in tooltip or in side windows, on demand.

3.2. 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.

3.2.1. Conventional dictionary

The conventional dictionary database allows the system to work in conventional dictionary mode. It is also used to build indexes for predictive typing, suffix-search, rhyme search etc.

3.2.2. WordNet

WordNet [4] is a large lexical database, consisting of synonym sets of words - “synsets” – structured by part-of-speech and numerous types of semantic relations. The richness of its structural information makes it highly acceptable resources for various NLP tasks. [6]. In the proposed framework, it provides glosses to be browsed by the user, and semantic database for basic word sense disambiguation.

3.2.3. BalkaNet, EuroWordNet, etc.

EuroWordNet [7] is a multilingual set of semantic databases for European languages, which is aligned to WordNet and to each other. It consists of databases for Dutch, Italian, Spanish, French, German, Czech and Estonian

BalkaNet [1] 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, Bulgarian BalkaNet is used. However, the system could easily be extended with other databases from BalkaNet or EuroWordNet frameworks, thus making it possible for “Smarty” to operate as English-Greek, English-Romanian, English-Serbian etc. comprehension assistant.

3.3. Natural language processing stages

3.3.1. POS-tagging

Selecting a word in a 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 like English, this could narrow the set of returned dictionary entries two or three times.

3.3.2. Lemmatisation and normalisation

Lemmatisation is the process of finding the basic form of a token.

This stage saves the user the trimming of words copied from text, thus speeds up the look-up.

3.3.3. Multiword expressions recognition

At this stage the context of the selected word is checked for possible multiword expressions.
The words from the context are lemmatised and then fuzzy-matched to patterns from a multi-word expressions database. Different techniques are applied to compute the degree of match: bag of words, POS-matching, regular expressions. The expressions with the highest degrees of match are presented to the user.

3.3.4. Word sense disambiguation

The ultimate goal of comprehension assistants is to find the most appropriate translation in given context. Smarty benefits from WordNet and aligned to it lexical databases for word sense disambiguation, because this approach allows the precise sense in the target languages to be found directly, using the alignment between the two databases. Word sense disambiguation contributes to further narrowing down 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 word sense disambiguation.

4. Evaluation

Several evaluation experiments are currently being conducted and the final version of the paper, if accepted, will detail on results of these experiments. One experiment aims to establish the efficiency it terms of time saving when using Smarty as compared with conventional electronic dictionaries. Another evaluation scenario envisages the assessment of users' feedback/questionnaires on the usefulnees of Smarty. Finally, an extrinstic evaluation experiment seeks to identify the impact of Smarty in the process of language learning.

References

1. BalkaNet - Design and Development of a Multilingual Balkan WordNet. http://www.ceid.upatras.gr/Balkanet

2. BulNet - Bulgarian WordNet. http://dcl.bas.bg/BulNet/general_en.html

3. Feldweg, Helmut and Breidt, Elisabeth (1996). "COMPASS An Intelligent Dictionary System for Reading Text in a Foreign Language". In Kiefer, F. and G. Kiss, editors, Papers in Computational Lexicography, COMPLEX '96, Budapest, pages 53-62.

4. Princeton Univeristy WordNet 2.0 Database documentation - wnstats.7WN.html.

5. Prószéky, Gábor (2002), "Comprehension Assistance Meets Machine Translation". In: Toma3 Erjavec; Jerneja Gros (eds) Language Technologies, 1–5. Institut Jo3ef Stefan, Ljubljana, Slovenia.

6. Mitkov, R.(2003), The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press.

7. EuroWordNet Results and Exploitation, http://www.illc.uva.nl/EuroWordNet/results-ewn.html

Appendix: Figures

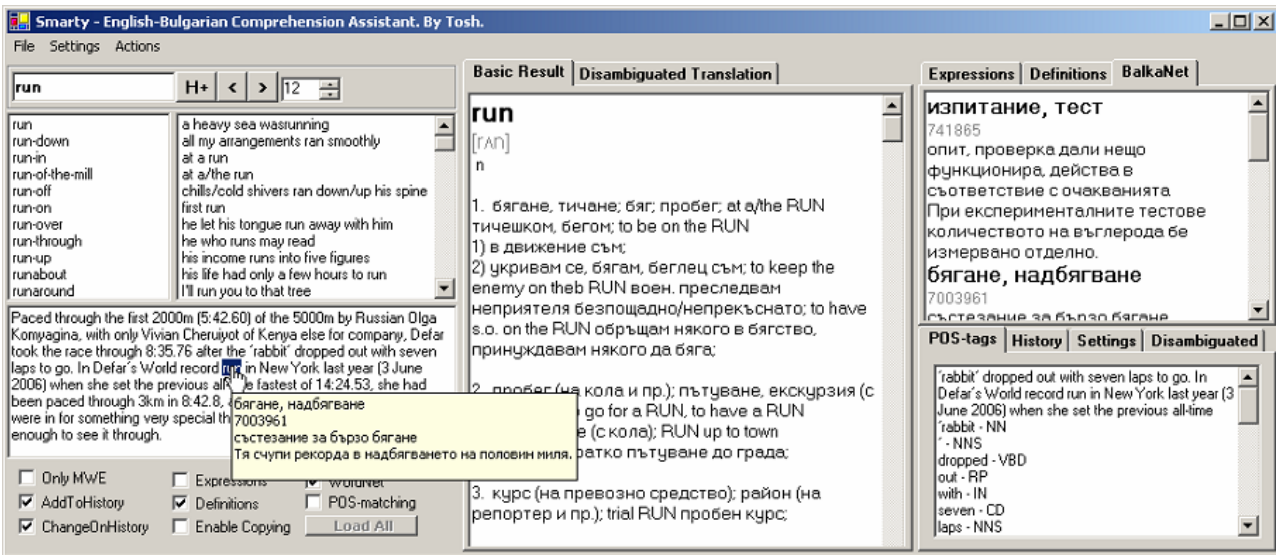

Figure 1

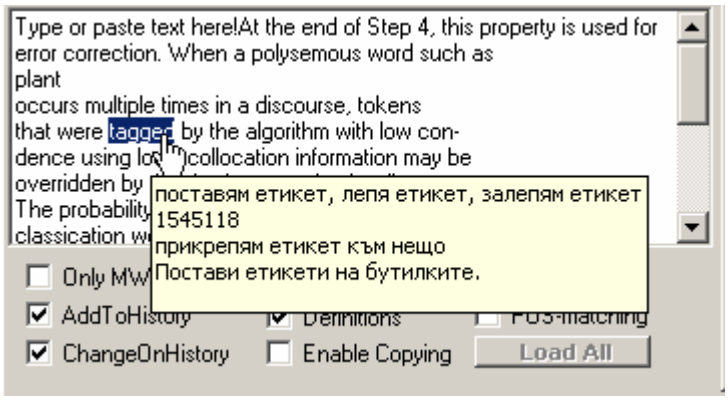

Figure 2

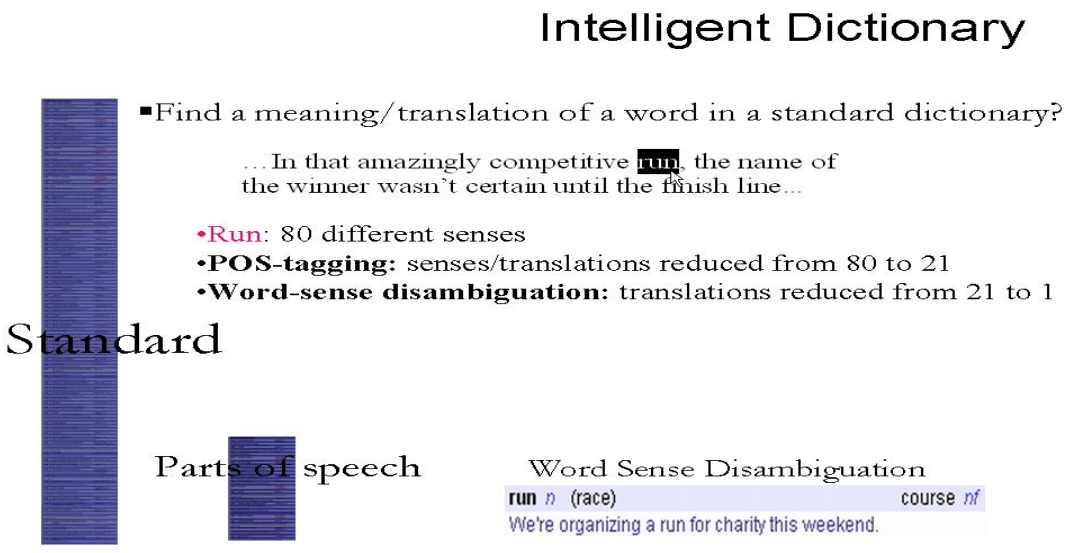

Figure 3