Browsing Multilingual Information with the MultiSemCor Web Interface

Marcello Ranieri, Emanuele Pianta, Luisa Bentivogli

ITC-irst
Via Sommarive 18, 38050 Povo (Trento) - Italy
{ranieri,pianta,bentivo}@itc.it

Abstract

Parallel and comparable corpora represent a crucial resource for different Natural Language Processing tasks like machine translation, lexical acquisition, and knowledge structuring but are also suitable to be consulted by humans for different purposes, such as linguistic teaching, corpus linguistics, translation studies, lexicography, multilingual information browsing. To enhance their exploitation by human users, specially designed interfaces need to be developed. In this paper we present the design and implementation of the MultiSemCor Web Interface. MultiSemCor is a parallel English/Italian corpus, which is being developed at ITC-irst starting from the English corpus SemCor. In MultiSemCor the texts are aligned at word level and semantically annotated with WordNet senses. The MultiSemCor Web Interface allows the users to exploit at best the potentiality of the corpus. We will describe the main functions of the interface, which provides two distinct browsing modalities: a bi-text-oriented modality and a word-oriented modality, which amounts to a bilingual semantic concordancer. Moreover, the MultiSemCor Web Interface is integrated with the on-line MultiWordNet browser, which gives access to the reference lexicon for MultSemCor.

1 Introduction

In the last years, the importance of parallel and comparable corpora has become more and more evident within the human language technology field, where these resources are used for the extraction of multilingual information in many tasks such as machine and machine-aided translation, linguistic teaching, lexicography, and knowledge structuring.

To enhance the exploitation of parallel corpora by humans, suitable interfaces need to be developed. Such interfaces should give access to all the information available in the corpus in an easy and intuitive way, and should possibly be integrated with other linguistic resources such as on-line dictionaries.

In this paper we will focus on the design and implementation of the MultiSemCor Web interface. MultiSemCor (Bentivogli & Pianta, 2002) is a parallel English-Italian corpus, aligned at word level and annotated with PoS, lemma and word sense. It has been obtained starting from SemCor, an English corpus semantically tagged with WordNet senses.

The rest of the paper is organized as follows. In Section 2 we summarise the methodology developed for the creation of the MultiSemCor corpus and its composition up to now. In Section 3 we describe in detail the MultiSemCor Web interface, its main browsing functionalities and novel characteristics. In Section 4 we outline some existing related work before concluding in Section 5.

2 The MultiSemCor Corpus

MultiSemCor is a parallel English-Italian corpus, which is being developed at ITC-irst starting from SemCor, a subset of the English Brown corpus containing almost 700,000 running words. In SemCor all the words are tagged by PoS, and more than 200,000 content words are also lemmatized and sense-tagged according to WordNet (Fellbaum, 1998). The strategy for creating MultiSemCor consists in having SemCor texts translated into Italian by professional translators; aligning Italian and English texts at word level; and then transferring the word sense annotations from English to the aligned Italian words. Both the word alignment and the annotation transfer are carried out automatically.

The main hypothesis underlying this methodology is that, given a text and its translation into another language, the translation preserves to a large extent the meaning of the source language text. A pilot study estimated that this methodology can be applied with a precision of 95% and a recall of 75%. The automatic projection of annotations from one language to another has been adopted as a strategy aiming at reducing the effort needed for obtaining annotated corpora (Pianta & Bentivogli, 2003): the result is an Italian corpus annotated with PoS, lemma and word sense, but also an aligned parallel corpus lexically annotated with a shared inventory of word senses. More specifically, the sense inventory used is MultiWordNet (Pianta et al., 2002), a multilingual lexical database in which the Italian component is strictly aligned with the English Princeton Wordnet.

At present MultiSemCor is composed of 116 English texts aligned at sentence level with their corresponding 116 Italian translations. The total amount of running words is 230,738 for English and 233,178 for Italian. The word alignment and transfer methodology has been applied to 29 texts out of the 116 texts available. These 29 texts are aligned at word level and annotated with PoS, lemma, and word sense. As regards English, we have 55,935 running words and 29,655 words semantically annotated (from SemCor). As for Italian, the corpus is composed of 59,726 running words among which 23,095 words are annotated with word senses that have been automatically transferred from English.

MultiSemCor will be useful for a variety of tasks. From a computational point of view we are planning to use it to automatically enrich the Italian component of MultiWordNet. As a matter of fact, out of the 23,095 Italian words automatically sense-tagged, 5,292 are not yet present in MultiWordNet and will be added to it. Moreover, MultiSemCor is also suitable to be consulted by humans for different purposes, such as language teaching and learning, translation studies, lexicography, multilingual information browsing.

3 The Interface

To help human users exploiting at best the potentiality of MultiSemCor, a Web-based browser has been realized. In its design we faced a number of interesting issues, such as making available to the users information about corpus annotation, bilingual text alignment, bilingual semantic concordancing, integration between corpora and lexical resources. To meet all these requirements, two distinct browsing modalities have been implemented. The first is *text-oriented* and the second is *word-oriented*. Each of these two modalities is embodied in a dynamic Web page.

3.1 Bi-text Browsing

In the text-oriented browsing modality, for each bi-text the user can access the following information:

A. alignment at sentence level
B. alignment at word level
C. dictionary of all the tokens of the text, with links to the sentences in which they occur

These functionalities have been implemented through a web-page organized in three sections corresponding to the three kinds of information above, see Figure 1. Section A contains the whole bi-text and shows the alignment at sentence level. This has been realized through a simple two column table, where each column contains the text in one of the two languages, and each row shows the alignment between a sentence and its translation. This solution shows the alignment between sentences, while keeping the possibility for the user to read the entire two texts in a natural way.

Section B allows the user to focus on a specific sentence and shows the available alignments at word level for that sentence. Showing word level alignments through a Web interface, while keeping the readability of the sentence in which the words occur is not as straightforward as showing sentence level alignments. Alignments could be shown for instance by marking aligned words with various colours, a colour for each alignment, or by putting the two sentences in a two column table, where each row contains a word alignment. However, we think that the former solution may be visually awkward, and for long sentences it makes the correspondence between words hard to trace. The latter solution makes the correspondence between words easier to read, but makes the entire sentence difficult or impossible to read, because of the vertical layout of words, and because the order of words in the target sentence needs to be completely changed. To solve the problem we choose to show only one word alignment per time, by highlighting the aligned words in the source and target sentence. Note that along with the word alignment, Section B also provides the available morphosyntactic information about the aligned words.

Section C of the interface contains a list of all the tokens in the current text in alphabetic order, with the translation in the other language. In fact there are two such lists, one for English-to-Italian, and one for Italian-to-English correspondences. Each token is hyperlinked with the sentence in which the token occurs.

In the example in Figure 1, the user is browsing the text br-c02 in Section A of the interface. By clicking on the word *character* contained in sentence nr. 73, he/she gets two results. Section B highlights the alignment between the word *character* in the English sentence, and *carattere* in the Italian translation. On the other hand, the top of Section C shows all the translations of the word *character* in the current text. Note that the user can now ask for the interface to show the passages in which the other translations of *character* are to be found.

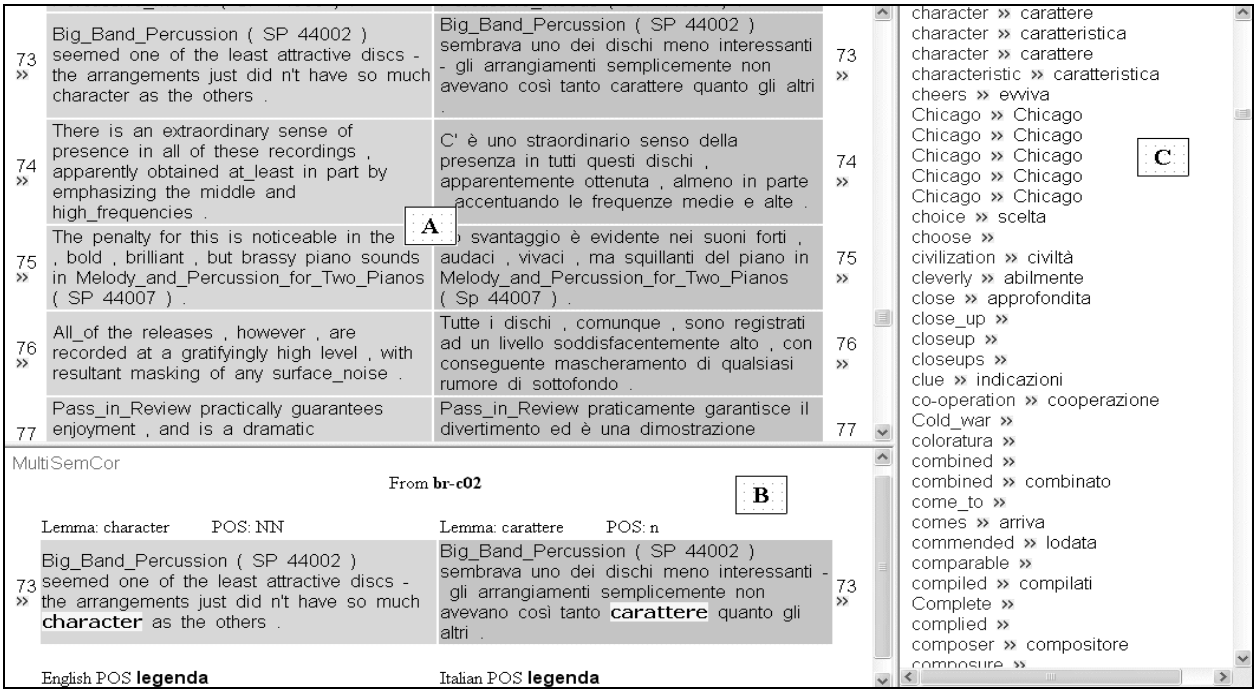

Figure 1: the browser in the *text-oriented* modality

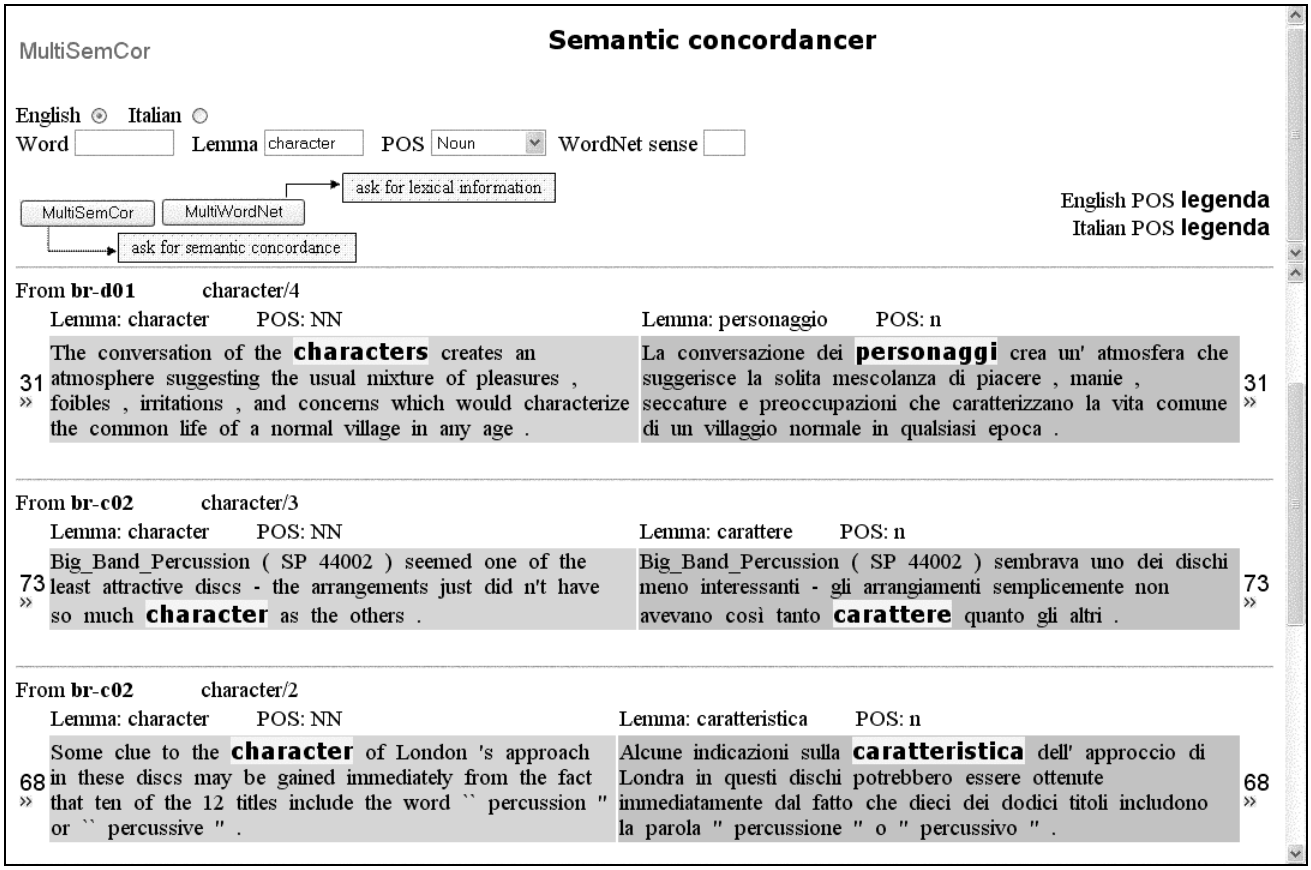

Figure 2: the result of a query in the semantic concordancer

3.2 Semantic Concordancer

The second modality for browsing the corpus is word-oriented, and amounts to a bilingual semantic concordancer, that is a tool able to provide all the occurrences of a certain word sense in a corpus. More precisely, in the MultiWordNet concordancer the user can alternatively search for all the occurrences of a *word form*, *lemma*, or *word sense* (according to MultiWordNet). The user can also constrain the search to a certain PoS. Free combinations between all these constraints (language, word form, lemma, word sense, PoS) are allowed. For instance the user can search for all the occurrences of: the word form *characters*; or the word form *character* as verb; or the lemma *character* in all of its senses; or the lemma *character* in its third sense (according to MultiWordNet).

The system will return a KWIC-like concordance of all the tokens in the corpus that match the request, within the sentence in which they occur; each sentence is presented along with its translation. Morphosyntactic information and the WordNet sense are also reported, as shown in Figure 2. An hyperlink connects each semantic concordance to the text-oriented browser, so that the user can easily get the bi-text in which a certain sentence occurs.

In Figure 2, the user has asked for the semantic concordance of the lemma *character* as a noun. Three aligned sentence in which the lemma occurs can be seen in the picture. Note that both singular and plural forms of the lemma have been selected, and the various senses of the word *character* (nr. 4, 3, and 2 with reference to MultiWordNet) are all translated with different Italian words.

3.3 Integration with MultiWordNet

Another important characteristic of the MultiSemCor Web interface is that it allows for the integration between the semantically annotated corpus and its reference lexicon, i.e. MultiWordNet.

This integration has a twofold effect. On the one side, while browsing the MultiSemCor word senses the user can consult MultiWordNet for a better understanding of the semantic annotation. On the other side, while browsing MultiWordNet the user can get examples of usage of a certain word sense from MultiSemCor. To our knowledge, MultiSemCor is the first interface to a multilingual corpus integrated with an on-line lexical resource.

The same form used in Figure 2 to ask for a semantic concordance, can be exploited to access the MultiWordNet lexical information related to a word form or lemma. See the “MultiWordNet” button next to the “MultiSemCor” button in the picture above. Figure 3 shows the result of searching lexical information about the lemma *character* in the standard MultWordNet interface. The two circles in the picture highlight two special icons. Clicking on one of them amounts to activating the MultiSemCor semantic concordancer on the specific sense which is in the focus of the interface.

From an implementation point of view, the MultiSemCor browser has been developed in PHP. The MultiSemCor corpus is encoded according to the XCES guidelines and it is stored in a MySQL database.

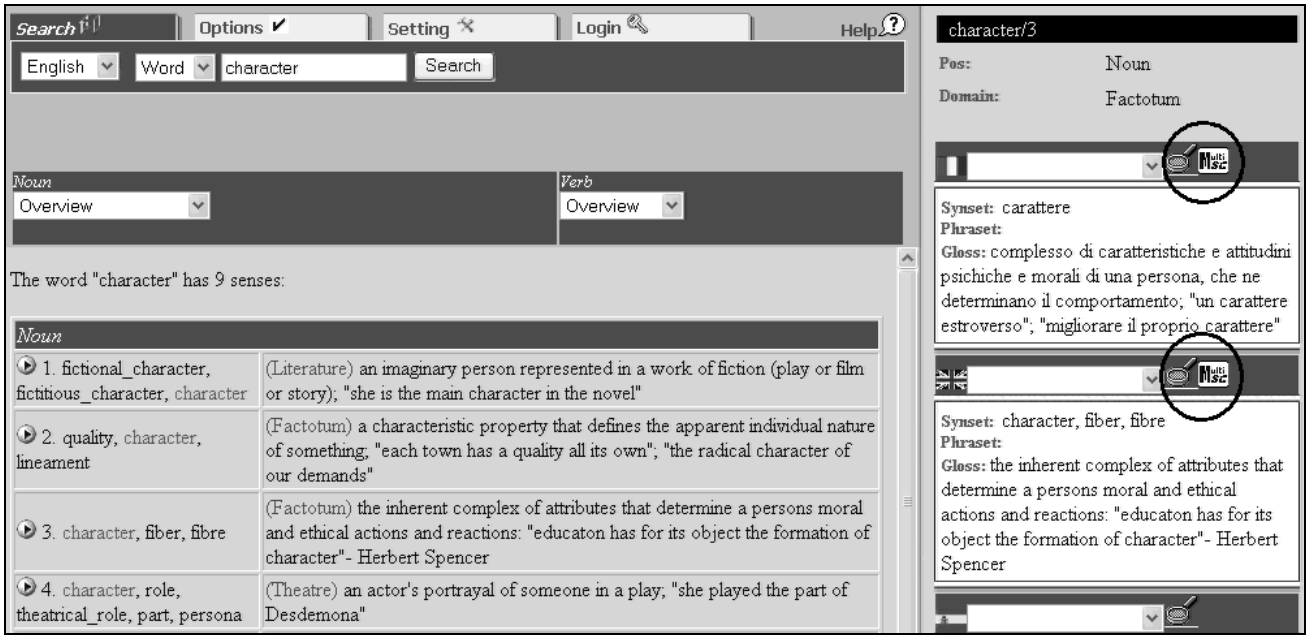

Figure 3: The MultiWordNet browser

4 Related Work

A number of institutions are active to collect, promote, and make available mono- and multilingual language resources and tools. The most important institutions, such as the LDC, ELRA/ELDA, Tractor, UCREL, and RALI, all distribute parallel or multilingual corpora. Also some parallel concordancers have been made available to the community. The most well known are: MultiConcord, ParaConc, WordSmith Tools, Web Concordancer, and TransSearch. Also, a number of projects built parallel corpora, and made them available through a Web interface. The project that is most similar to MultiSemCor is the multilingual English-Catalan-Castillan parallel corpus, developed at Universitat Pompeu Fabre of Barcelona. See http://terminotica.upf.es/academic/. This is the only available interface giving access to word-level alignment. Other on-line interfaces allow for the browsing of sentence-level alignment, and for a token-based search in the text:

- the bilingual English-Chinese parallel corpus by the Hong Kong Virtual Language Center: http://www.edict.com.hk/concordance
- the bilingual English-Portuguese parallel corpus Compara, by the Linguateca group: http://www.linguateca.pt/COMPARA
- the bilingual English-Slovene parallel corpus by the University of Ljubljana-Slovenia: http://nl2.ijs.si/index-bi.html

Other projects made available only an on-line sample. These are the Web TCE interface to the bilingual English-Norwegian parallel corpus at the University of Oslo, and the TransSearch interface to the Canadian Hansard Corpus.

5 Conclusion

In this paper we presented an on-line, freely accessible Web interface to MultiSemCor, a parallel English/Italian corpus, annotated at lexical level. The interface gives access to a large amount of bilingual information through two main modalities, addressing the needs of users with different background. Moreover, it allows for the integrated access to the MultiWordNet on-line lexical database. A first version of the on-line MultiSemCor browser is available at the following address: http://tcc.itc.it/projects/multisemcor.

References

Bentivogli, L., & Pianta, E. (2002). Opportunistic Semantic Tagging. In Proceedings of the Third International Conference on Language Resources and Evaluation (pp. 1401--1406). Las Palmas, Canary Islands – Spain, May 29-31, 2002.

Fellbaum, C. (ed.) (1998). Wordnet: An Electronic Lexical Database. Cambridge (Mass): The MIT Press.

Pianta, E., & Bentivogli, L. (2003). Translation as Annotation. In Proceedings of the AI*IA 2003 Workshop "Topics and Perspectives of Natural Language Processing in Italy"(pp. 40--48). Pisa, Italy, September 2003.

Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In Proceedings of the 1st International Global WordNet Conference (pp. 293--302), Mysore, India, January 21-25, 2002.

MultiWordNet, http://tcc.itc.it/projects/multiwordnet/multiwordnet.php