Multilingual Word Sense Disambiguation and Entity Linking for Everybody

Andrea Moro, Francesco Cecconi, and Roberto Navigli

Sapienza University of Rome, Viale Regina Elena 295, 00198, Italy
{moro,cecconi,navigli}@di.uniroma1.it

Abstract. In this paper we present a Web interface and a RESTful API for our state-of-the-art multilingual word sense disambiguation and entity linking system. The Web interface has been developed, on the one hand, to be user-friendly for non-specialized users, who can thus easily obtain a first grasp on complex linguistic problems such as the ambiguity of words and entity mentions and, on the other hand, to provide a showcase for researchers from other fields interested in the multilingual disambiguation task. Moreover, our RESTful API enables an easy integration, within a Java framework, of state-of-the-art language technologies. Both the Web interface and the RESTful API are available at http://babelfy.org

Keywords: Multilinguality, Word Sense Disambiguation, Entity Linking, Web interface, RESTful API

1 Introduction

The tasks of Word Sense Disambiguation (WSD) and Entity Linking (EL) are well-known in the computational linguistics community. WSD [9, 10] is a historical task aimed at assigning meanings to single-word and multi-word occurrences within text, while the aim of EL [3, 12] is to discover mentions of entities within a text and to link them to the most suitable entry in the considered knowledge base. These two tasks are key to many problems in Artificial Intelligence and especially to Machine Reading (MR) [6], i.e., the problem of automatic, unsupervised understanding of text. Moreover, the recent upsurge of interest in the use of semi-structured resources to create novel repositories of knowledge [5] has opened up new opportunities for wide-coverage, general-purpose Natural Language Understanding techniques. The next logical step, from the point of view of Machine Reading, is to link natural language text to the aforementioned resources.

In this paper, we present a Web interface and a Java RESTful API for our state-of-the-art approach to WSD and EL in arbitrary languages: Babelfy [8]. Babelfy is the first approach which explicitly aims at performing both multilingual WSD and EL at the same time. The approach is knowledge-based and exploits semantic relations between word meanings and named entities from BabelNet [11], a multilingual semantic network which provides lexicalizations and glosses for more than 9 million concepts and named entities in 50 languages.

2 BabelNet

In our work we use the BabelNet 2.5[1] semantic network [11] since it is the largest available multilingual knowledge base and is obtained from the automatic seamless integration of Wikipedia[2], WikiData[3], OmegaWiki[4], WordNet [4], Open Multilingual WordNet [1] and Wiktionary[5]. It is available in different formats, such as via its Java API, a SPARQL endpoint and a linked data interface [2]. It contains more than 9 million concepts and named entities, 50 million lexicalizations and around 250 million semantic relations (see http://babelnet.org/stats for more detailed statistics). Moreover, by using this resource we can leverage the multilingual lexicalizations of the concepts and entities it contains to perform disambiguation in any of the 50 languages covered in BabelNet.

3 The Babelfy System

Our state-of-the-art approach, Babelfy [8], is based on a loose identification of candidate meanings (substring matching instead of exact matching) coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations. Here we briefly describe its three main steps:

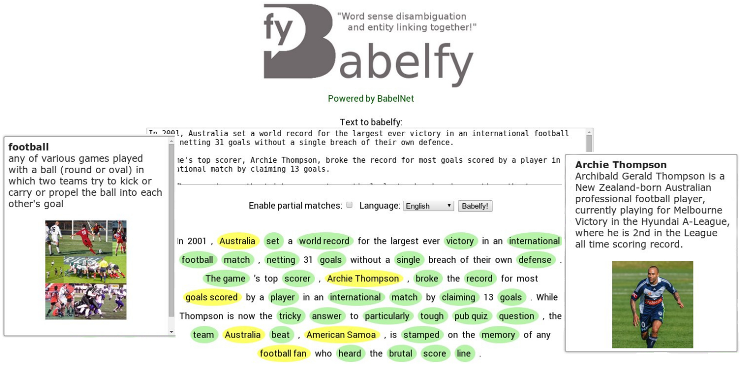
1. Each vertex, i.e., either concept or named entity, is automatically associated with a semantic signature, that is, a set of related vertices by means of random walks with restart on the BabelNet network.
2. Then, given an input text, all the linkable fragments, i.e., pieces of text being equal to or substring of at least one lexicalization contained in BabelNet, are selected and, for each of them, the possible meanings are listed according to the semantic network.
3. A graph-based semantic interpretation of the whole text is produced by linking the candidate meanings of the selected fragments using the previously-computed semantic signatures. Then a densest subgraph heuristic is used to extract the most coherent interpretation and finally the fragments are disambiguated by using a centrality measure within this graph.

A detailed description and evaluations of the approach are given in [7, 8].

4 Web Interface and RESTful API

We developed a Web interface and a RESTful API by following the KISS principle, i.e., "keep it simple, stupid". As can be seen from the screenshot in Figure

[1] http://babelnet.org
[2] http://www.wikipedia.org
[3] http://wikidata.org
[4] http://omegawiki.org
[5] http://wiktionary.org


"Word sense disambiguation
and entity linking together!"
fy Babelfy
Powered by BabelNet
Text to babelfy:
Enable partial matches: □
Language: English ▾
Babelfy!

In 2001, Australia set a world record for the largest ever victory in an international football
netting 31 goals without a single breach of their own defence.

ne's top scorer, Archie Thompson, broke the record for most goals scored by a player in
ational match by claiming 13 goals.

In 2001 , Australia set a world record for the largest ever victory in an international football match , netting 31 goals without a single breach of their own defense . The game 's top scorer , Archie Thompson , broke the record for most goals scored by a player in an international match by claiming 13 goals . While Thompson is now the tricky answer to particularly tough pub quiz question , the team Australia beat , American Samoa , is stamped on the memory of any football fan who heard the brutal score line .

football
any of various games played with a ball (round or oval) in which two teams try to kick or carry or propel the ball into each other's goal

Archie Thompson
Archibald Gerald Thompson is a New Zealand-born Australian professional football player, currently playing for Melbourne Victory in the Hyundai A-League, where he is 2nd in the League all time scoring record.

Fig. 1. A screenshot of the Babelfy Web interface.

1, the Web interface asks for the input text, its language and whether the partial matching heuristic should be used instead of the exact string matching one. After clicking on "Babelfy!" the user is presented with the annotated text where we denote with green circles the concepts and with yellow circles the named entities. As for the Java RESTful API, users can exploit our approach by writing less than 10 lines of code. Here we show a complete example:

```
// get an instance of the Babelfy RESTful API manager
Babelfy bfy = Babelfy.getInstance(AccessType.ONLINE);
// the string to be disambiguated
String inputText = "hello world, I'm a computer scientist";
// the actual disambiguation call
Annotation annotations = bfy.babelfy("", inputText,
    Matching.EXACT, Language.EN);
// printing the result
for(BabelSynsetAnchor annotation : annotations.getAnnotations())
    System.out.println(annotation.getAnchorText()+"\t"+
        annotation.getBabelSynset().getId()+"\t"+
        annotation.getBabelSynset());
```

4.1 Documentation for the RESTful API

```
Annotation babelfy(String key, String inputText,
    Matching candidateSelectionMode, Language language)
```

The first parameter is the access key. A random or empty key will grant 100 requests per day (but a less restrictive key can be requested). The second parameter is a string representing the input text (sentences or whole documents can be input up to a maximum of 3500 characters). The third parameter is an enum with two possible values: EXACT or PARTIAL, to enable, respectively, the exact or partial matching heuristic for the selection of fragment candidates found in the input text. The fourth parameter is the language of the input text (among 50 languages denoted with their ISO 639-1 uppercase code).

`Annotation` is the object that contains the output of our system. A user can access the POS-tagged input text with getText() which returns a list of WordLemmaTag objects with the respective getters. With getAnnotations() a user will get a list of BabelSynsetAnchor objects, i.e., the actual annotations. A user can use getAnchorText() to get the disambiguated fragment of text and with getBabelSynset() get the selected Babel synset. Moreover, if a user wants to anchor the disambiguated entry to the input text, the start and end indices of the tagged text can be gotten with getStart() and getEnd().

5 Conclusion

In this paper, we presented and described the typical use of the Web interface and Java RESTful API of our state-of-the-art system for multilingual Word Sense Disambiguation and Entity Linking, i.e., Babelfy, available at http://babelfy.org

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

References

1. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. In: Proc. of ACL. pp. 1352–1362 (2013)
2. Ehrmann, M., Cecconi, F., Vannella, D., Mccrae, J.P., Cimiano, P., Navigli, R.: Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In: Proc. of LREC. pp. 401–408 (2014)
3. Erbs, N., Zesch, T., Gurevych, I.: Link Discovery: A Comprehensive Analysis. In: Proc. of ICSC. pp. 83–86 (2011)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
5. Hovy, E.H., Navigli, R., Ponzetto, S.P.: Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence 194, 2–27 (2013)
6. Mitchell, T.M.: Reading the Web: A Breakthrough Goal for AI. AI Magazine (2005)
7. Moro, A., Navigli, R., Tucci, F.M., Passonneau, R.J.: Annotating the MASC Corpus with BabelNet. Proc. of LREC pp. 4214–4219 (2014)
8. Moro, A., Raganato, A., Navigli, R.: Entity Linking meets Word Sense Disambiguation: A Unified Approach. TACL 2, 231–244 (2014)
9. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2), 1–69 (2009)
10. Navigli, R.: A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In: Proc. of SOFSEM. pp. 115–129 (2012)
11. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
12. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115 (2013)