Enhancing the Accuracy of Ancient Greek WordNet by Multilingual Distributional Semantics

Yuri Bizzoni, Riccardo del Gratta, Federico Boschetti, Marianne Reboul

Enhancing the Accuracy of Ancient Greek WordNet by Multilingual Distributional Semantics

Yuri Bizzoni[1], Riccardo Del Gratta[1], Federico Boschetti[1], Marianne Reboul[2]
[1]ILC-CNR, Pisa
[2]Université de Paris 4, Paris
{yuri.bizzoni,riccardo.delgratta}@gmail.com,
federico.boschetti@ilc.cnr.it, marianne.reboul@free.fr

Abstract

English. We discuss a method to enhance the accuracy of a subset of the Ancient Greek WordNet based on the Homeric lexicon and the related conceptual network, by using multilingual semantic spaces built from aligned corpora.

Italiano. *Esponiamo un metodo per migliorare l'accuratezza di un sottoinsieme dell' Ancient Greek WordNet, basato sul lessico Omerico e sulla relativa rete concettuale, attraverso l'uso di spazi semantici plurilingui costruiti su corpora paralleli allineati.*

1 Introduction

The Ancient Greek WordNet (AGWN) represents the first attempt to build a WordNet for Ancient Greek (Bizzoni et al., 2014).

The AGWN synsets are aligned to Princeton WordNet (PWN) (Fellbaum, 1998), to Italian WordNet (IWN) (Roventini et al., 2003), developed at the Institute for Computational Linguistic "A. Zampolli" in Pisa, to the Italian section of MultiWordNet,[1] developed at Bruno Kessler Foundation and to a Latin WordNet (LWN) created with the same criteria of AGWN and linked to Minozzi's Latin WordNet (Minozzi, 2009) and (McGillivray, 2010), developed at the University of Verona. In this way the user is allowed to find the equivalents of a set of synonyms into different languages. The AGWN can be freely accessed through a Web interface,[2] which allows enabled users to add or delete words in the synsets, adapt the glosses and validate the lexico-semantic relations.[3] We first created AGWN by bootstrapping Greek-English pairs from bilingual dictionaries and by assigning Greek words to PWN synsets associated to the corresponding English translations. As a drawback of this method, a large number of synsets and lexico-semantic relations are spuriously over-generated by English homonymy and polysemy. As exposed in (Bizzoni et al., 2014), to have PWN as a pivoting resource[4] propagates the same drawback to other connected WordNet in CoPhiWordNet Platform. In order to improve the accuracy of a subset of AGWN synsets related to the Homeric lexicon and the related conceptual network, we have automatically extracted word translations from Greek-Italian parallel texts by applying distributional semantic strategies illustrated in the following sections and verified how many of these translation were in CoPhiWn. According to the methodology explained in (Francis Bond and Uchimoto, 2008), trilingual resources (in our case the original Greek-English pairs extracted from dictionaries and the Greek-Italian pairs extracted from aligned translations) are useful to enhance the accuracy of a bootstrapped WordNets.

2 Translation Mining through Semantic Spaces

We present a way to automatically improve the accuracy of Ancient Greek word translations by applying the principles of distributi-

[1]http://multiwordnet.fbk.eu
[2]GUI beta-version at http://www.languagelibrary.eu/new_ewnui

[3]In the following, when we use the term CoPhiWordNet Platform (CoPhiWn) we mean the three WordNets: AGWN, IWNand PWN.
[4]For example,PWN links through ILI (Vossen, 1998) AGWN to IWN

onal semantics to aligned corpora (Dumais et al., 1997) and (Yuri, 2015). We will first explain the ratio of this method and then show how it is useful to improve AGWN in several ways (see Section 2.7). Although Ancient Greek obviously does not have native speakers, we dispose of a great variety of translations of the same classical texts written in several languages and different historical periods. The study of large diachronical corpora of translations is both relevant in classical studies and a valuable source of information to build or improve the accuracy of multilingual lexico-semantic resources (see Section 3).

2.1 Aligning long and literary-biased translations to the original text

We applied a strategy to automatically align Greek-Italian parallel corpora through two main steps: in the first step we segmented texts in small portions; in the second step we linked those texts together. The result is that each Ancient Greek segment is aligned to its translations. After the segment-to-segment alignment, we applied the distributional semantics method illustrated below, in order to identify word-to-word translations.

2.2 Distributional Semantics

It is argued by several linguists (Miller, 1971) and (Firth, 1975) that one of the best ways to define the meaning of a word is the study of the relations with the other words in the close context. So it is possible to hypothesize that we learn the meaning of many new words thanks to the way they are linked to words we already know, and in general, that we learn the meaning of words by perceiving their verbal as well as non-verbal context. We can study semantic similarities between terms by quantifying their distribution: similar words will have similar contexts. In the same way, we can suppose that, in an aligned parallel corpus, a word and its translation will tend to appear in the same aligned segments. For this reason, the contextual segment of the original Greek word and the contextual aligned segment of the translation have the same identifier.

2.3 Semantic Spaces based on aligned corpora

There are several kinds of linguistic contexts that can be selected to study word similarity (Lenci, 2008):

- window-based collocates: two words co-occur if they appear in a given context window;
- text regions: two words co-occur if they appear in a same textual area such as a document, a paragraph, and so on;
- syntactic collocates: two words co-occur if they appear in a same syntactic pattern, for example if they are the direct objects of a verb, etc.

Although the most typical approach to distributional semantics is the use of window-based collocates, this kind of context becomes useless in multilingual corpora, since words in different languages do not share a common context. We use the method based on text regions collocates, which considers every couple of aligned segments as the default textual area. Word vectors of 0s and 1s in both languages are constructed accordingly to the absence/presence of the word in the aligned couple.
Thus, Ancient Greek and Italian words are mingled together in the vectorial space.[5]

2.4 Words and their translations tend to be neighbors

With a similar procedure, Ancient Greek and Italian equivalent words will happen to have similar vectors, since they will appear in the same aligned chunks. Consequently they will be close in the resulting semantic space. To compute the proximity of vectors we used the cosine similarity measure (Sahlgren, 2006).

2.5 Parts of Speech TRanslations

Performance on nouns is higher than performance on verbs, adjectives and adverbs, due to larger translational fluctuations for the latter parts of speech. Anyway, although verbs

[5] In our experiment the resulting vector has a dimension of $\sim 60k$

are more polysemous than nouns, we apparently are able to find relevant verb translations: *uccidere - kteíno (to kill)*, *morire - thnésko* (to die), *amare - philéo (to love)* and even *essere - eimí (to be)*. The same holds for adjectives, but, however,we found acceptable results also in this category: *bello - kalòs (beautiful)*, *nobile - agauòs (noble)*. Interestingly, from color adjectives we were only able to retrieve black and white translations: *neromélas (black)*, *bianco-leukós (white)*. Color adjectives in Ancient Greek are naturally complex to analyze, since it is hard to retrieve their exact meaning in absence of speakers; this indetermination apparently propagates to our outcomes.

Finally, it is also relevant to observe that extremely polysemous categories like adverbs in some cases find a correct translation: *ek - fuori (out)*, *non-ou - non (not)*.

2.6 Data Presentation and Some Results

We extracted the five most similar items for 121 Ancient Greek words (randomly chosen from different groups of frequency) from a semantic space built on the original texts, i.e five complete Iliad translations and four complete Odyssey translation in Italian aligned to the original texts. The original data resulted in 605 rows (121 time 5pairs); when it comes to verify whether a Greek/Italian pair is mapped in CoPhiWn, we expect that the modern polysemy, the one inducted by English to Italian mapping will increase the number of pairs. Indeed, we found that 605 pairs correspond to 736 Greek-English-Italian possible triples. However, only 176 triples have been successfully found in CoPhiWn. A manual validation of the resulting set excluded 13 triples which are caused by the modern polysemy reducing the found triples to 164. Not surprisingly, the coverage of the triples in CoPhiWn $\sim 23\%$ is quite close to the coverage of AGWN, cf. (Bizzoni et al., 2014) ($\sim 28\%$).

2.7 AGWN: strenghtening bilingual links

If an Ancient Greek word is linked to an Italian word in CoPhiWn and it is distributionally near to the same Italian word in a semantic space, the probability that this link is correct is high. For instance, the word *pólemos*, frequent in Homer, is linked in CoPhiWn to an Italian synset composed by the words *guerra, battaglia, ostilità*. The first two terms appear also to be the nearest Italian terms to the word *pólemos (war, battle)* in our semantic space. This match helps us to increase the probability that *guerra* and *battaglia* are sound translations of *pólemos*, and thus that the Italian and Greek synsets are correctly interlinked.

In CoPhiWn the word *hémar (day)* is linked to the synonyms *giorno, giornata*, and in our semantic space it appears very similar to the word *giorno* only. But the distributional information from our semantic space reinforces the association between *hémar* and the overall Italian synset.

This way to retrieve crosslingual information from textual corpora is highly helpful to discover errors due to the employ of polysemy in different languages. For instance, in CoPhiWn, the word *astér (star)* is linked both to the synset associated to the word *stella*, glossed as star in the sky and to the synset associated to the word *divo*, glossed as star in the show business, due to the intermediation of the English word *star*,[6] while, as expected, *astér* is distributionally similar only to *stella* in our semantic space. The word *dóru (spear and mast)* is linked on one hand to *asta, arma* synset and on the other hand to *prora, prua*, glossed as parts of the boat, which is synecdochically related to the mast, but in our semantic space it appears near only to the words of the first group, allowing us to score higher only the first equivalence. It is important to remember that we can incur in cases of stylistically biased translations and synonyms: *árma (charriot)* can be *cocchio* or *carro* in different translations.

Additional examples are the following: the most similar terms to Italian *mare* in our semantic space are *thálassa, háls, póntos*, three words indicating the concept of sea clustered together by their common translation. *scudo* (shield) is associated both to *aspís* and *sakós*, *soffio* (breath) leads to *pnóe* and *ánemos*, through *popolo* (people) we find *láos, démos* and among the most similar words of *dolore* (pain) we find both *pénthos* and *álgos*. With the same mechanisms that allow to find word to word

[6] This is one effect of the modern polysemy described in Section 2.6.

translations, we can find also some small sets of potential synonyms in the same language looking at their distributional behavior: so *aithér* is near to *oúranos* and *hétor* is near to *thumós*.

2.8 CoPiWn! (CoPiWn!): supporting hypernym/hyponym relations

A system based on distributional semantics tends to cluster together not only bilingual synonyms and translations, but also hypernyms and hyponyms. They tend to have distributionally similar, although not identical, behaviors, and it can easily happen that a word is translated with a hypernym, or more rarely with a hyponym, in another language. Systems to discriminate between hypernyms and synonyms in semantic spaces could become very useful in this context. See for example (Benotto, 2013) and Lenci et al. 2012.

3 Conclusions and Future Work

We have elaborated a system to enhance the accuracy of Ancient Greek WordNet. This system appears to be useful to verify the soundness of automatically generated links between the Ancient Greek WordNet and WordNet in other languages. The method aims at increasing the precision of the Greek-Italian pairs within their translations, since it removes modern polysemy and discards translations in CoPhiWn that are not supported by actual texts' translations.

References

Giulia Benotto. 2013. Modelli distribuzionali delle relazioni semantiche: il caso dell'iperonimia. *Animali, Umani, Macchine. Atti del convegno 2012 del CODISCO*.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta , Monica Monachini, and Gregory Crane. 2014. The Making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, MA, USA.

John Rupert Firth. 1975. *Modes of meaning*. College Division of Bobbs-Merrill Company.

Kyoko Kanzaki Francis Bond, Hitoshi Isahara and Kiyotaka Uchimoto. 2008. Boot-strapping a wordnet using multiple existing wordnets. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, 20(1):1–31.

Barbara McGillivray. 2010. Automatic selectional preference acquisition for Latin verbs. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pages 73–78. ACL.

George A Miller. 1971. Empirical methods in the study of semantics. *Semantics, an interdisciplinary reader in philosophy, linguistics, and psychology*, pages 569–585.

Stefano Minozzi. 2009. The Latin WordNet Project. In Peter Anreiter and Manfred Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137 of *Innsbrucker Beiträge zur Sprachwissenschaft*, pages 707–716.

Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Christian Girardi, Bernardo Magnini, Rita Marinelli, and Antonio Zampolli. 2003. Italwordnet: building a large semantic database for the automatic treatment of Italian. *Computational Linguistics in Pisa, Special Issue, XVIII-XIX, Pisa-Roma, IEPI*, 2:745–791.

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Bizzoni Yuri. 2015. The Italian Homer - The Evolutions of Translation Patterns between the XVIII and the XXI century. Master's thesis, University of Pisa.