International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:8, 2015

Contextual Distribution for Textual Alignment

Yuri Bizzoni, Marianne Reboul

Abstract**—**Our program compares French and Italian translations of Homer's *Odyssey,* from the XVIth to the XXth century. We focus on the third point, showing how distributional semantics systems can be used both to improve alignment between different French translations as well as between the Greek text and a French translation. Although we focus on French examples, the techniques we display are completely language independent.

Keywords**—**Translation studies, machine translation, computational linguistics, distributional semantics.

I. Introduction

WE compare French and Italian translations of Homer's *Odyssey*, from the XVIth to the XXth century. Open data algorithms are still either too dependent on language specifications and databases or unreliable. We hope to overcome these aporias. The Greek text is first cut on anchor points (proper nouns), and so is its corresponding translation; the corpus is then aligned with our algorithm and divided in fixed chunks. Each Greek chunk is given a fixed ID, allowing us to give its translations the corresponding IDs. Each translation is therefore aligned one to another according to their identification.

The alignment of the source to the target is done in three steps (preprocessing, alignment and postprocessing). To align textual chunks we use three main systems: 1, an automatically generated bilingual dictionary of Greek-French proper nouns; 2, length and frequency measures; 3, a dictionary of distributionally related terms.

II. Distributional Semantics

The third point allowed us to consider a token not just as one data unit but as a contextual vector.

A problem in aligning different monolingual translations is that different translators could use different words to express the same meaning, and it would be necessary to find a way to detect the semantic similarity between their different choises. A way to model the semantic similarity of two elements is to study the problem from a distributional point of view, which is done through the construction of contextual vectors.

A contextual vector represents the distributional behaviour of a word in a corpus. The distribution of a word is the list of contexts in which such word appears [1], and it gives a representation of how that word is used [2].

It is argued by several linguists [2], [3] that one of the best ways to define the meaning of a word is to look at that word in relation to others.

Yuri Bizzoni is graduated in Computational Linguistics at the University of Pisa (e-mail is yuri.bizzoni@gmail.com).

Marianne Reboul is a PhD student at the University La Sorbonne of Paris (e-mail: odysseuspolymetis2010@gmail.com).

The way two words are used can be considered as an indication of their difference in meaning [4]: thus, words with similar distributions should have similar meanings. Words having similar contextual vectors will probably share a similarity in meaning: they could be synonyms, since they are used in the same contexts.

III. Monolingual Distributional Similarities

Comparing vectors in both source and target allowed us to determine a distributional dictionary of potential synonyms.

We saw, in fact, that contextual features could still be useful in different translations to determine synonymy.

Some contexts tend to remain similar from the source to the target, and therefore may be most useful for chunk-to-chunk or even word-to-word alignment. Just to make an example, we can look at the following lines taken respectively from Dacier's and Sommer's translations:

une hécatombe de taureaux et d'*agneaux*
(Dacier, *Odyssée*, I)

une hécatombe de taureaux et de *brebis*
(Sommer, *Odyssée*, I)

In this example, *agneaux* and *brebis* have exactly the same context, thus it is possible to hypothesize a semantic similarity between the two words.

Although stylistic differences between translators involve large changes also in lexicon, it is often the case that two different synonyms, or pseudo-synonyms, are used in similar contexts, allowing us to distributionally detect similar variations. To do so, we give each word of each text (stored in a non repetitive map) a modifiable immediate context.

The choice of the context has a central role in this model, since it strongly conditions the results. For example, a 4-word contextual window will take into account the two words preceding and the two words following every occurrence of the given term:

la ville **sacrée** de Troie (Dacier, Odyssée, I)

les murs **sacrés** de Troie (Sommer, Odyssée, I)

From the preceding example, it is already possible to induce that *sacrée* and *sacrés* have some distributional similarity, since they share at least a part of context *(de Troie*). With different window sizes, this information could be reinforced by new elements, or lost in noise. Some researchers set a reduced co-occurrence window of 4 or 5 words, while others prefer larger ones, of the order of 100 words [4]. We chose a 4-word window.

In the next step, a word vector can be created defining the co-occurrence of the word with every other term in the text.

This way, it is possible to represent the semantic similarity of two words as the similarity between their vectors. Co-occurrence vectors are set into a co-occurrence matrix. Such matrix normally has a set of words in rows and a set of words in columns while cells contain the frequency of co-occurrence of each word in rows with each word in columns:

TABLE I
CO-OCCURRENCE MATRIX

	la	ville	les	murs	**de**	**Troie**	...
sacrés	0	0	1	1	**1**	**1**	
sacrée	1	1	0	0	**1**	**1**	

A co-occurrence matrix is a semantic space. A semantic space is a multidimensional model of word distribution in a text or corpus, having as many dimensions as the distributional vectors and as many points as the number of words. Therefore, each word is stored as a vector of contextual co-ocurrences. Sahlgren [5] explains that such a model of word distribution allows a useful *similarity-is-proximity* metaphor: words with similar vectors represent points with proximal locations. The locations of the words in the semantic space do not reveal much about their meaning or their use. It is the *relative* location of words which matters (the fact that a word A is nearer to a word B than to a word C). In a semantic space, it is not important to know where a word is but rather how distant it is from another word.

When all the distributional vectors are ready, we can measure their relative proximity with the cosine similarity.

This similarity metric takes the scalar product of two vectors and divides it by the product of their norms:

$$\text{sim}\cos(\vec{x},\vec{y}) = \frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

This is useful because it overcomes the frequency issue: by normalizing the scalar product of two vectors, the effects their length may cause are neutralized, simply because longer vectors (vectors with larger values) will also have higher norms. It also gives a fixed similarity measure: two identical vectors will have a cosine similarity of 1 and two orthogonal vectors will have a cosine similarity of 0. Using the cosine similarity, the length of the vectors does not matter.

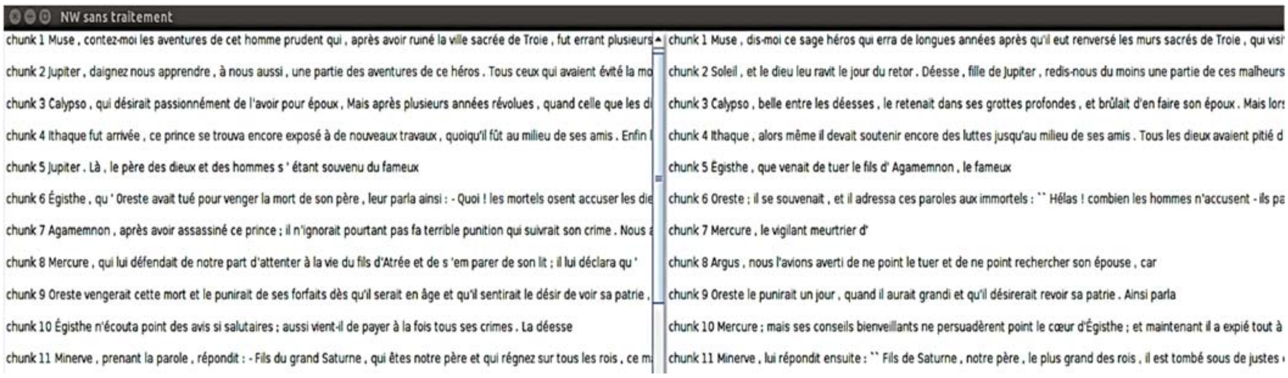

Fig. 1 Needleman-Wunsch alignment without contextual semantic distribution

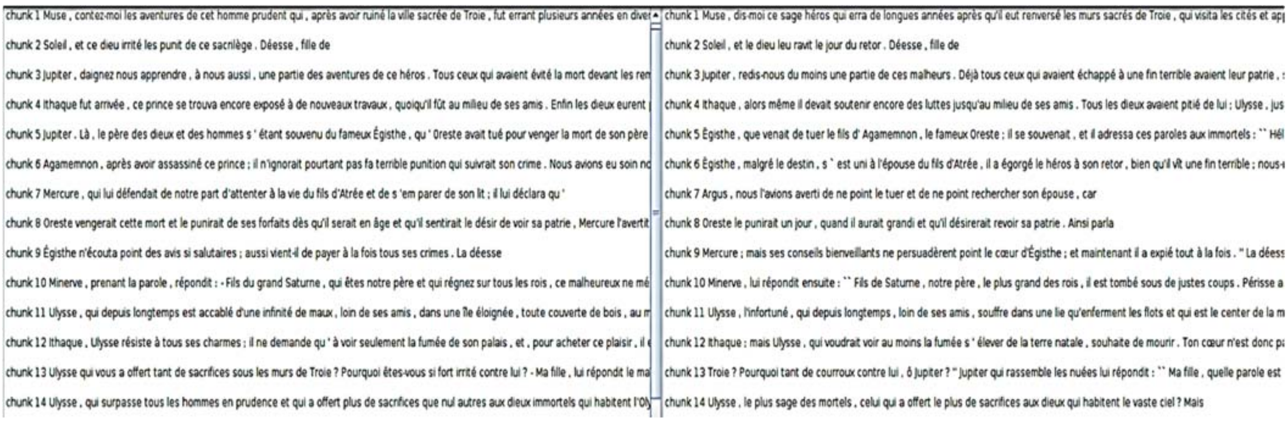

Fig. 2 Needleman-Wunsch alignment with fixed contextual semantic distribution

If the cosine similarity result is high, we store each word and its potential proximity tokens in a distributional dictionary that will impact on the final similarity score.

Referring to the preceding example, a chunk with *agneaux* and a chunk with *brebis* will have a slightly higher probability to be aligned - thus, to contain the same information - than two chunks with words distributionally unrelated.

The immediate results show that distributionally near words tend to be either semantically related or linked by similar expressions, and in general that this technique allows us to improve the alignment of translational segments.

In Fig. 1, we can see that, although some chunks have been correctly aligned, many mistakes remain. For 17 chunks, 7 are faulty. In Fig. 2 however, when context is taken into account, only 3 mistakes remain (which could be reduced to one, as two of these problematic alignments are to be considered in reverse).

The theoretical interest of these results in our line of work is also to be considered: the changing in the use and the meaning of words is of primary interest in translation studies. The same words could have very different distributional neighbours in different translations. The fact that contextual information can be succesfully used to infer semantic similarities between translations of different eras can be fascinating to consider.

This method being entirely language independent, it may be adaptable to any monolingual set of translation.

Once the preprocessing is done, an adaptation of Needleman-Wunsch's algorithm (initially created to align protein sequences) [6] associates each chunk in a potentially final aligned corpus.

This algorithm works building up a grid from any two sequences. For each element in the first sequence (for example, for each letter, or for each segment) it assigns a value of matching probability to every element of the second sequence, based on a given similarity score and on the already made matches.

The similarity score is calculated through a specific function that uses some pre-defined metric to determine how much two elements are similar between them. This is somehow the most sensitive part of the system, since it is the function that decides whether two elements have a good probability of matching. The function that attributes a similarity score determines the success of the rest of the operation. In our case, since we are using non-annotated corpora, we maintained very simple parameters: the similarity is calculated through the automatically generated dictionary and some other heuristics.

We use the distributional similarity between words to improve the precision of the similarity score.

IV. Cross-Lingual Distributional Similarities

Naturally, a context-based similarity is very helpful between monolingual translations.

Vectorial representations are widely used in linguistics to model the distance between words, concepts [7], expressions [8], etc., butsemantic distance is normally computed between two words of the same language and only recently some studies have been made about vectors in bilingual parallel corpora. Corpus-based approaches to parallel corpora have been exploited mainly in the field of Machine Translation. Cohn and Lapata [9] try to improve poor-resource languages translation through a triangulation method, using a rich language as pivot between two texts. Banea [10] uses multilingual corpus-based approach to improve sentiment analysis annotation. In general, standard context-based distributional analysis is bound to work only on monolingual texts.

Thus, to embetter Greek-French alignment we used a slightly different technique, that can be applied in a second-round alignment to refine results.

In this case, two aligned parts of a bilingual text can be considered as a unique cooccurrence window, or, better, as a unique "word area" that can, or cannot, contain some given words in both languages.

In this perspective, the vector of each word of the parallel corpus (thus, the vector of every word independently from the language it belongs to) is determined by the presence or absence of that word in each bilingually aligned block. Being the blocks composed of a segment of text in a language and its equivalent in the other language, we could expect from an absolutely literal translation to return perfectly similar vectors for each word and its translation.

So, from a first alignment we obtain Greek-French coupled chunks and we build our words' vectors looking at whether each word appears or not in a determined Greek-French couple. Ancient Greek and French equivalent words will happen to have similar vectors, since they will appear in the same aligned chunks.

The principle is simple: we create a semantic space of the word-to-document kind, so that in rows are words and in columns are textual blocks in which those words can appear. Each textual block is composed by two parallel segments already aligned. One word's vector is given by its presence or absence in textual blocks. Consequently, both Greek and French words can appear in every block - can have a non-zero value in every position of their vector.

A Greek word and its French rendition will tendentially have very similar vectors and thus will appear very near, as in the following toy-example:

Ulysse sur les vaisseaux recourbés vers Ilion
ὈδυσσῆοςἼλιον εἰς εὔπωλον ἔβη κοίλῃσ' ἐνὶ νηυσίν

Cyclope tua dans sa caverne profonde
Κύκλωψ ἐν σπῆϊ γλαφυρῷ

le fils chéri d'**Ulysse**
Ὀδυσσῆος φίλος υἱός

Ὀδυσσῆος vector: **1 0 1**
Ulysse vector: **1 0 1**
Cyclope vector: 0 1 0

This system, a form of cross-lingual term-by-document matrix, is already known in information retrieval although it is mainly used to retrieve documents, and not single terms, in different languages. Basically, a query in a language is used to find relevant documents in another language.

This technique can both allow a word-to-word research on text and give better alignment results when connected to the aligner, since it gives a quick way to find new anchor words for the text. Starting from a broad block-to-block alignment with the heuristics we described, it is possible to reach a more refined matching through the extraction of single word translations, that can be used in a second round alignment as additional anchor words.

From this basic idea an improved dictionary of anchor words can be created, with values of probability assigned to each Greek-French translation, and a second-round alignment can be run to obtain more accurate results.

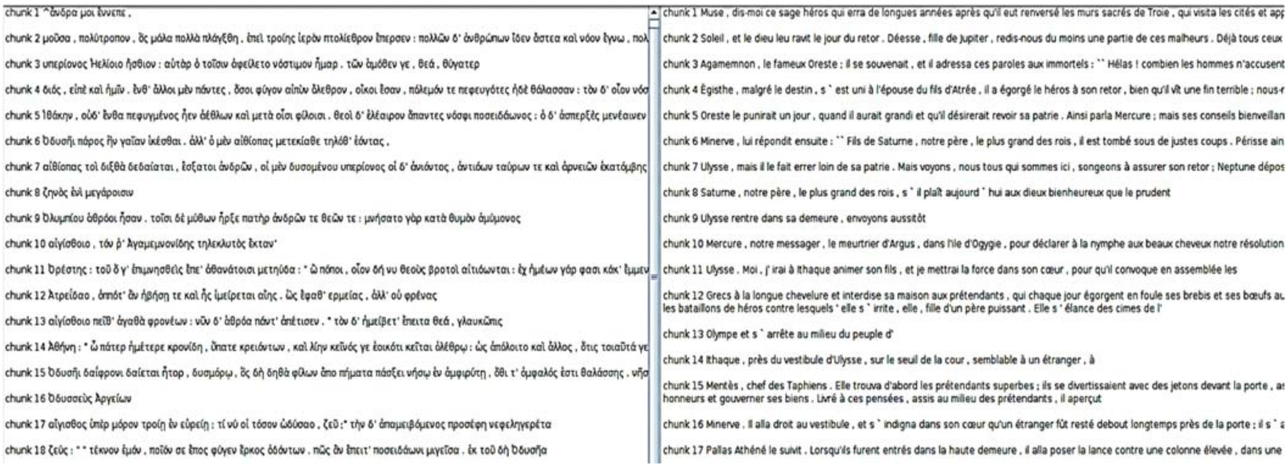

Fig. 3 Needleman-Wunsch Greek alignment without distributional semantics

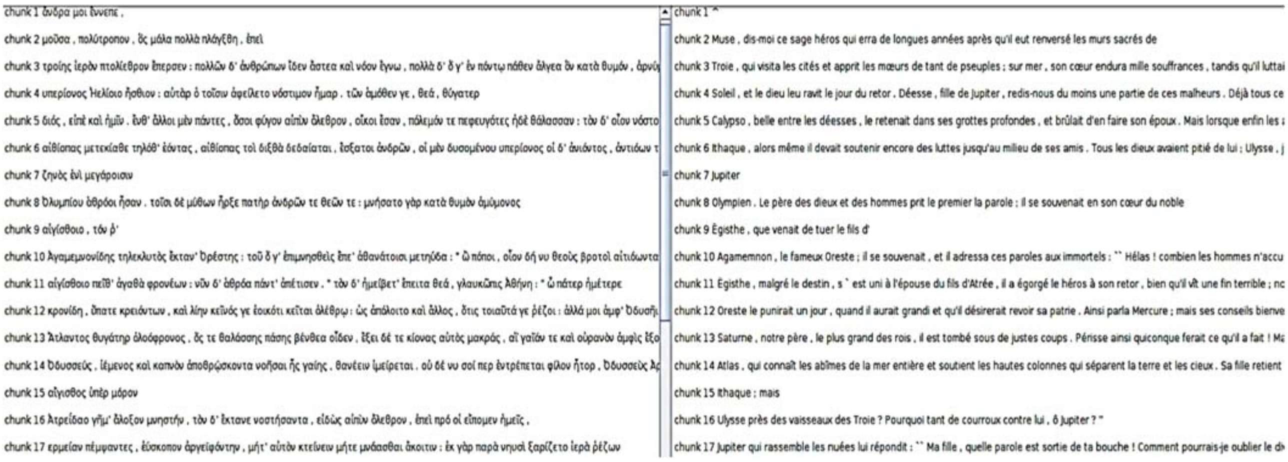

Fig. 4 Needleman-Wunsch Greek alignment with distributional semantics

In Fig. 3 we can see that many chunks are not correctly aligned. At least 9 of the 17 chunks have not found their correct match. However, in Fig. 4, considering the post-processing of pre-segmented distributional semantics, the result is almost perfect: 3 out of 17 chunks have found their correct match. It is therefore visible that this ultimate step, based on realigning preceding chunks and applying distributional semantics methods for a last alignment, is most effective.

V. Conclusion

As a language may be defined as a system based on grammatical principles (which may be flexible or not), any language may not be organized totally arbitrarily. Words and their multiple meanings are defined and clarified by their context. Therefore, understanding the logic behind a simple multi-character token implies a deep consideration of not only the word examined, but also of the whole group of words that surrounds it. This theoretical principle may also be applied on a statistical point of view: even in texts made to be impossible to understand, language has its logic, and words cannot be considered independently. Thus, in a statistical approach, if we may not strictly speaking infer the meaning of words on the sole consideration that they may be similar, we can at least conclude that each word cannot be considered as a nucleus, but as a particle of a much more complex cell. As a result, we have shown that alignment procedures need not only to consider a word through its internal similarity with others, but also as a necessary part of a larger statistical system. Studying context for alignment is an image of the way the human brain works: understanding a language means understanding its systematic principles.

REFERENCES

[1] Miller, G. A. (1967), *Empirical methods in the study of semantics*, in Journeys in Science: Small Steps – Great Strides, University of New Mexico Press, Albuquerque: 51–7.
[2] Miller G. and Charles W. (1991), *Contextual correlates of semantic similarity*. In: Language and Cognitive Processes 6(1):1–28.
[3] Firth J. R. (1951), *Modes of meaning*, In: Essays and Studies, The English Association, Oxford.
[4] Harris, Z., (1954), *Distributional structure*. Word, X/2-3, pp. 146-62.
[5] Sahlgren, M., (2006). *The Word Space Model*. PhD Dissertation, Stockholm University.
[6] Needleman, S. and Wunsch, C., (1970), *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. In: Molecular Biology, n.48, pp.443-453.
[7] Alfonseca,E. and Manandhar,S., (2002), *Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures*. Berlin: Springver-Veralg.
[8] Baroni,M., Bernardi,R., Do,N., Shan,C., (2012), *Entailment above the word level in distributional semantics*. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp.23-32.
[9] Pado S. and Lapata M. (2007). *Dependency-based Construction of Semantic Space Models*. Computational Linguistics, 33:2, 161-199.
[10] Banea C., Mihalcea J. R., and Wiebe J., (2010). *Multilingual subjectivity: Are more languages better*? In: Proceedings of COLING'10.