Exploring a Continuous and Flexible Representation of the Lexicon

Pierre Marchal, Thierry Poibeau

Exploring a Continuous and Flexible Representation of the Lexicon

Pierre Marchal
ERTIM, INaLCO
2 rue de Lille
F-75007 PARIS
pierre.marchal@inalco.fr

Thierry Poibeau
LaTTiCe, CNRS – ENS – Université Paris III
PSL – USPC
1 rue Maurice Arnoux
F-92120 MONTROUGE
thierry.poibeau@ens.fr

Abstract

We aim at showing that lexical descriptions based on multifactorial and continuous models can be used by linguists and lexicographers (and not only by machines) so long as they are provided with a way to efficiently navigate data collections. We propose to demonstrate such a system.

1 Background and Motivations

"You shall know a word by the company it keeps!" (Firth, 1957). This all too well-known citation motivates any lexicographic work today: it is widely accepted that word description cannot be achieved without the analysis of a large number of contexts extracted from real corpora. However, this is not enough.

The recent success of deep learning approaches has shown that discrete representations of the lexicon are no longer appropriate. Continuous models offer a better representation of word meaning, because they encode intuitively valid and cognitively plausible principles: semantic similarity is relative, context-sensitive and depends on multiple-cue integration.

At this point, one may say that it doesn't matter if these models are too abstract and too complex for humans as they are used by machines. We think this argument is wrong. If continuous models offer a better representation of the lexicon, we must conceive new lexical databases that are usable by humans and have the same basis as these continuous models. There are arguments to support this view.

For example, it has been demonstrated that semantic categories have fuzzy boundaries and thus the number of word meanings per lexical item is to a large extent arbitrary (Tuggy, 1993). Although this still fuels lots of discussions among linguists and lexicographers, we think that a description can be more or less fine-grained while maintaining accuracy and validity. Moreover, it has been demonstrated that lexical entries in traditional dictionaries overlap and different word meanings can be associated with a sole example (Erk and McCarthy, 2009), showing that meaning cannot be sliced into separate and exclusive word senses.

The same problem also arises when it comes to differentiating between arguments and adjuncts. As said by Manning (2003): "There are some very clear arguments (normally, subjects and objects), and some very clear adjuncts (of time and 'outer' location), but also a lot of stuff in the middle". A proper representation thus need to be based on some kind of continuity and should take into consideration not only the subject and the object, but also the prepositional phrases as well as the wider context.

Some applications already address some of the needs of lexicographers in the era of big data, *i.e.* big corpora in this context. The most well-known application is the SketchEngine (Kilgarriff et al., 2014). This tool has already provided invaluable services to lexicographers and linguists. It gives access to a synthetic view of the different usages of words in context. For example, the SketchEngine can give a direct view of all the subjects or complements of a verb, ranked by frequency or sorted according to various parameters. By exploding the representation, this tool provides an interesting view of the lexicon. However, in our opinion, it falls short when it comes to showing the continuous nature of meaning.

Here we propose a system that combines the advantages of existing tools (a wide coverage database offering a synthetic view of a large vocabulary) with those of a dynamic representation. We focus on verbs since these lexical items offer the most complex syntactic and semantic behaviors. More specifically, we examine Japanese verbs, as Japanese is a language that presents a complex system of case markers that are generally semantically ambiguous.

2 Outline of our approach

When building a verb lexicon, numerous challenges arise such as the notion of lexical item – that is, how many entries and subentries are necessary to describe the different meanings of a given verb? – and the distinction between arguments and adjuncts – that is, what complements are necessary to describe a particular meaning of a given verb? Following up on studies in natural language processing and linguistics, we embrace the hypothesis of a continuum between ambiguity and vagueness (Tuggy, 1993), and the hypothesis that there is no clear distinction between arguments and adjuncts (Manning, 2003). Although this approach has been applied and evaluated for Japanese, the theoretical framework to compute the argumenthood of a complement, or build the hierarchical structure of the lexical entries, is partially independent.

We assume a list of verbal structures that have been automatically extracted from a large representative corpus. A verbal structure is an occurrence of a verb and its complements (expressed as syntactic dependencies); a complement is an ordered pair of a lexical head and a case marker.

Computing the argumenthood of complements Following up on previous studies on the distinction between arguments and adjuncts (Manning, 2003; Merlo and Esteve Ferrer, 2006; Fabre and Bourigault, 2008; Abend and Rappoport, 2010), we propose a new measure of the degree of argumenthood of complements, derived from the famous TF-IDF weighting scheme used in information retrieval:

$$\text{argumenthood}(v, c) = (1 + \log \text{count}(v, c)) \log \frac{|V|}{|\{v' \in V : \exists(v', c)\}|} \quad (1)$$

where c is a complement (*i.e.* an ordered pair of a lexical head and a case particle); v is a verb; $\text{count}(v, c)$ is the number of cooccurrences of the complement c with the verb v; $|V|$ is the total number of unique verbs; $|\{v' \in V : \exists(v', c)\}|$ is the number of unique verbs cooccurring with this complement. That is, we are dealing with complements instead of terms, and with verbs instead of documents. This measurement captures two important rules of thumb for distinguishing between arguments and adjuncts. The first part of the formula $(1 + \log \text{count}(v, c))$ takes the idea that complements appearing frequently with a given verb tend to be arguments; the second part of the formula $\log \frac{|V|}{|\{v' \in V : \exists(v', c)\}|}$, that complements which appear with a large variety of verbs tend to be adjuncts.

The proposed measure assigns a value between 0 and 1 to a complement – 0 corresponds to a prototypical adjunct; 1 corresponds to a prototypical argument – and thus model a continuum between arguments and adjuncts.

Enriching verb description using shallow clustering A verbal structure corresponds to a specific sense of a given verb; that is, the sense of the verb is given by the complements selected by the verb. Yet a single verbal structure contains a very limited number of complements. So as to obtain a more complete description of the verb sense, we propose to merge verbal structures corresponding to the same meaning of a given verb into a minimal predicate-frame using reliable lexical clues. We call this technique *shallow clustering*. Our method relies on the principles that *i)* two verbal structures describing the same verb and having at least one common complement might correspond to the same verb sense, and that *ii)* some complements are more informative than others for a given sense.

As for the second principle, the measure of argumenthood, introduced in the previous section, serves as a tool for identifying the complements which contribute the most to the verb meaning. Our method merges verbal structures in an iterative process – beginning with the most informative complements (*i.e.* complements yielding the highest argumenthood value) – as shown in Algorithm 1.

Data: A set $\mathbf{W}$ of verbal structures $(\mathbf{v}, \mathbf{D})$, where $\mathbf{v}$ is a verb and $\mathbf{D}$ is a list of complements
Result: A set $\mathbf{W'}$ of minimal predicate-frames $(\mathbf{v}, \mathbf{D'})$ such that $|W'| \leqslant |W|$
$W' \leftarrow \emptyset$;
foreach *verb* v **in** $\{v : \exists(v, D) \in W\}$ **do**
  `/* Let C' be the list of complements cooccurring with v sorted by argumenthood values in non-increasing order */`
  $C \leftarrow \{c : \exists(v, D) \in W \wedge c \in D\}$;
  $C' \leftarrow (c : c \in C \wedge$ `argumenthood(`$v, C'[i]$`)` $\geqslant$ `argumenthood(`$v, C'[i+1]$`)` $)$;
  for $i \leftarrow 0$ **to** `length(C')` $- 1$ **do**
    `/* Let D' be a subset of` $\{D : \exists(v, D) \in W\}$ `*/`
    $D' \leftarrow \emptyset$;
    foreach *list of complements* D **in** $\{D : \exists(v, D) \in W\}$ **do**
      if $C'[i] \in D$ **then**
        add D to D';
        remove (v, D) from W;
      end
    end
    foreach *list of complements* D **in** $\{D : \exists(v, D) \in W\}$ **do**
      if $\exists X \in D'$ *such that* $D \subset X$ **then**
        add D to D';
        remove (v, D) from W;
      end
    end
    if $|D'| \geqslant 2$ **then** add the minimal predicate-frame (v, D') to W';
  end
end

Algorithm 1: Shallow clustering of verbal structures.

Modeling word senses through hierarchical clustering We propose to cluster the minimal predicate-frames built during the *shallow clustering* procedure into a dendrogram structure. A dendrogram allows the definition of an arbitrary number of classes (using a threshold) and thus fits nicely with our goal of modeling a continuum between ambiguity and vagueness. A dendrogram is usually built using a hierarchical clustering algorithm, with a distance matrix as its input. So as to measure the distance between minimal predicate-frames, we propose to represent minimal predicate-frames as vectors which would then serve as arguments of a similarity function.

Following previous studies on semantic composition, we suppose that "the meaning of a whole is a function of the meaning of the parts and of the way they are syntactically combined" (Partee, 1995) as well as all the information involved in the composition process (Mitchell, 2011). The following equation summarizes the proposed model of semantic composition:

$$p = f(\mathbf{u}, \mathbf{v}, R, K) \tag{2}$$

where $\mathbf{u}$ and $\mathbf{v}$ are two lexical components; R is the syntactic information associated with $\mathbf{u}$ and $\mathbf{v}$; K is the information involved in the composition process.

Following the principles of distributional semantics (Harris, 1954; Firth, 1957), lexical heads can be represented in a vector space model (Salton et al., 1975). Case markers (or prepositions) can be used as syntactic information. Finally, we propose to use our argumenthood measure to initialize the K parameter as it reflects how important a complement is for a given verb.

The proposed model of semantic composition is applied recursively to all the complements of a given minimal predicate-frame so as to produce a single vector. Hierarchical clustering is then applied to vector

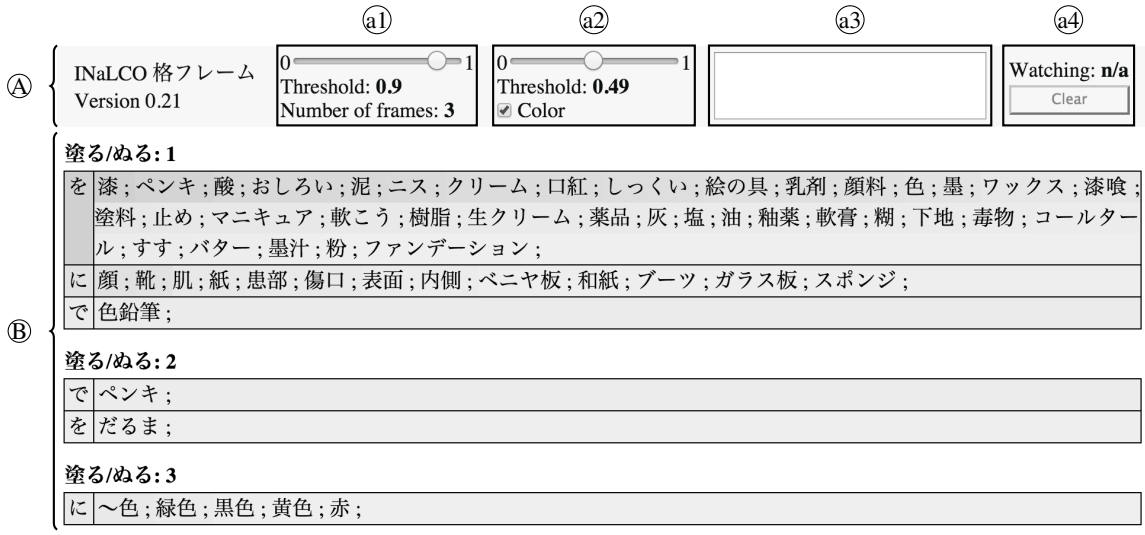

Figure 1: Screen capture of our visualization tool – Ⓐ control panel: ⓐ1 slider for partitioning subentries; ⓐ2 slider for selecting complements; ⓐ3 notification zone; ⓐ4 subentry identifier. – Ⓑ subentry panel. Here, we present the entry for the verb 塗る *nuru* "to smear". The threshold values reveal the locative alternation XにYを塗る *X ni Y wo nuru* "smear Y on X" ↔ XをYで塗る *X wo Y de nuru* "smear X with Y".

representation of the minimal predicate-frames so as to build a dendrogram for each verb in our data. The dendrogram serves as a model of the continuum between ambiguity and vagueness.

3 Overview of the visualization tool

In order to make the resource usable by humans, it is necessary to provide the end user with a graphical interface to navigate and explore the data in more detail. Our goal is to build a resource that reflects the subtleties of continuous models but avoids the complexity of a multifactorial analysis and offers a simple interface that allows a lexicographer or a linguist to navigate the data collection easily.

After many attempts, we managed to create a simple interface where the multifactorial analysis is abstracted as a double continuum: a continuum between ambiguity and vagueness, and a second continuum between arguments and adjuncts. Figure 1 shows a screen capture of our visualization tool[1].

Slider ⓐ1 represents the continuum between ambiguity and vagueness. It sets a threshold on the dendrogram of the subentries; subentries whose distance is less than the threshold are merged so as to make a single subentry. When the threshold is set to 0, each minimal predicate-frame corresponds to a distinct subentry; when set to 1 all minimal predicate-frames are merged into a single subentry. Slider ⓐ2 represents the continuum between arguments and adjuncts. It sets a threshold so as to only select complements that exhibit a certain argumenthood value. When the threshold is set to 0, all complements are displayed; when set to 1, only the complement with the highest degree of argumenthood is visible. Also, a color is assigned to each lexical head so as to indicate its degree of argumenthood: a light color indicates a value close to 0 (an adjunct); a dark color indicates a value close to 1 (an argument).

The user can move the two sliders back and forth to dynamically increase or decrease the number of subentries and complements. As the number of subentries can be substantial, we implemented various functionalities to track changes in the subentry panel. The notification panel ⓐ3 displays information about subentries that have merged or split, appeared or disappeared. We also implemented a mechanism to automatically focus and lock the subentry panel on a particular subentry (in which case the subentry number is given in ⓐ4).

[1]Our results are available online: `http://marchal.er-tim.fr/ikf`. Data is distributed under a Creative Commons licence (CC-BY-SA 4.0).

4 Discussion

First experiments with lexicographers have shown that exploration of the lexicon in the manner described above makes it possible to find new verb usages. The interface we have created is intuitive enough to allow the user to gradually unveil the meanings of verbs, starting with discriminative syntactic patterns (*e.g.* transitive versus intransitive) or broad semantic classes of complements (*e.g.* literal versus figurative), to finally uncover – as constraints on the partitioning of subentries and on the selection of complements are released – more fine-grained and domain-dependant meanings of the verbs. This exploration method also allows the user to observe linguistic phenomena at the syntax/semantics interface – such as diathesis alternations, as shown in Figure 1 with the locative alternation of the verb 塗る *nuru* "to smear" –, and verify prior assumptions that have been formulated in a different framework, particularly the status of certain complements (*i.e.* arguments versus adjuncts), or account for the productivity of some fixed expressions.

5 Conclusion

In this paper we have shown that is is possible to build lexical resources, based on continuous models, that can be useful not only to machines but also to humans. A more formal evaluation of both the interface and the lexical resource is currently underway, involving both linguists and lexicographers. This evaluation has already proved that our resource and its interface is useful, efficient and sufficiently powerful for professional end users.

Acknowledgements

Pierre Marchal's research has been partially funded by a "contrat doctoral" from the French Ministry of Higher Education and Research (Doctoral School 265). The authors wish to thank Ms. Jennifer Lewis-Wong (ERTIM, INaLCO) who assisted in the proof-reading of this paper.

References

Omri Abend and Ari Rappoport. 2010. Fully unsupervised core-adjunct argument classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 226–236.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449.

Cécile Fabre and Didier Bourigault. 2008. Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18(1):87–102.

John R. Firth, 1957. *A Synopsis of Linguistic Theory 1930–1955*, pages 1–32. Basil Blackwell, Oxford.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10:146–162.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Christopher D. Manning. 2003. Probabilistic syntax. In S. Jannedy R. Bod, J. Hay, editor, *Probabilistic Linguistics*, pages 289–341. MIT Press, Cambridge, MA.

Paola Merlo and Eva Esteve Ferrer. 2006. The notion of argument in prepositional phrase attachment. *Computational Linguistics*, 32(3):341–377.

Jeffrey Mitchell. 2011. *Composition in Distributional Models of Semantics*. Ph.D. thesis, University of Edinburgh.

Barbara H. Partee, 1995. *Lexical Semantics and Compositionality*, pages 311–360. The MIT Press, Cambridge, MA.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3):273–290.