Identification of Rare & Novel Senses Using Translations in a Parallel Corpus

Richard Schwarz*, Hinrich Schütze*, Fabienne Martin†, Achim Stein†

*Institute for Natural Language Processing, Universität Stuttgart
richard.schwarz@ims.uni-stuttgart.de
†Institut für Linguistik/Romanistik, Universität Stuttgart
{fabienne.martin,achim.stein}@ling.uni-stuttgart.de

Abstract

The identification of rare and novel senses is a challenge in lexicography. In this paper, we present a new method for finding such senses using a word aligned multilingual parallel corpus. We use the Europarl corpus and therein concentrate on French verbs. We represent each occurrence of a French verb as a high dimensional term vector. The dimensions of such a vector are the possible translations of the verb according to the underlying word alignment. The dimensions are weighted by a weighting scheme to adjust to the significance of any particular translation. After collecting these vectors we apply forms of the K-means algorithm on the resulting vector space to produce clusters of distinct senses, so that standard uses produce large homogeneous clusters while rare and novel uses appear in small or heterogeneous clusters. We show in a qualitative and quantitative evaluation that the method can successfully find rare and novel senses.

1. Introduction

The identification of rare and novel senses is a challenge in lexicography. We present a new method for finding such senses based on a *multitext*, a multilanguage parallel corpus. We concentrate on French verbs and show in a qualitative and quantitative evaluation that the method can successfully find rare and novel senses. The basic idea of our approach is to represent each occurrence of a French verb in the multitext as the signature of its translations. These signatures are then collected and clustered. Our expectation is that most resulting clusters represent standard uses, but that unusual and exceptional clusters, when analyzed manually, will yield rare and novel usages. Homogeneous clusters can be quickly identified after inspecting a few members. The main advantage of our method is that the lexicographer can concentrate on heterogeneous and exceptional clusters. Thus, we support rapid finding of interesting usages by allowing the user to discard the majority of standard uses quickly and efficiently.

2. Related Work

We start with the observation that polysemous words tend to be translated as distinct words in other languages. E.g., the two senses of English *sentence* are translated as *peine* and *phrase* in French. Early work exploiting this property of multitext includes (Brown et al., 1991), (Gale et al., 1993), and (Schütze, 1993). More recently, multitext approaches to word sense tagging have been proposed by Diab and Resnik (2002) and Ng et al. (2003). Related work on sense discrimination and analysis of senses that also used clustering includes (Ploux and Victorri, 1998), (Ide, 1999), (Ide et al., 2002), (Tufiş et al., 2004), and (Francois, 2007). Most of this work (with the exception of (Francois, 2007)) has been done for nouns. Alignment quality is typically higher for nouns than for verbs. There has also been work on trying to use clustering to improve the performance of machine translation systems (e.g., (Och, 1999) and (Uszkoreit and Brants, 2008)).

Our goal is not applied NLP, word sense tagging/disambiguation or elucidating distinctions between word senses based on dictionaries. Rather, we focus on speeding up the manual identification of rare and novel senses in support of lexicographers and curators of lexical databases.

3. Methodology

We use Europarl (Koehn, 2005) as our corpus, aligned with GIZA++ (Och and Ney, 2003). The Europarl corpus is a multilingual parallel corpus that has been extracted from the proceedings of the European Parliament. All nine Romanic and Germanic languages in the corpus have been used for this project, each ranging in size from 33 to 44 million tokens. Figure 1 shows an example of a GIZA++ word alignment for a French/English sentence pair. In the example the English *most* in the target sentence is aligned to the first four words of the source sentence *la plus grande partie* while *would* is aligned to *voudraient*.

3.1. Vector Space Model

Let T be the possible translations of a French verb v in Europarl. T is a set containing words from 8 languages. We represent an occurrence v_i of v as the T-dimensional vector $\vec{v_i}$ with a non-zero value for words aligned to v_i and 0 for all words not aligned to v_i. The exact values are determined by the term weighting we use, which is elaborated upon below. We call the set of vectors of all n occurrences of v the vector space V.

In practice we write the vector $\vec{v_i}$ as a list of the words that are actual translations of v_i and not bother with the ones that are not. As an example consider one occurrence v_i of *demander* in Europarl. It is represented as $\vec{v_i}$ = [*ask, verlangen, pedir, pretendere, willen, exige, kräver, forlanger*], based on the word alignment.

We try to account for the differences in significance of particular translations by applying term weighting. Some words have greater validity than others – e.g., the English word *whether* may occur as a translation of *demander* merely due to an alignment error. Let the alignment frequency $\mathrm{af}(v, t)$ be the number of times that verb v and translation t were aligned. In Europarl af(*demander, ask*)

```
la plus grande partie des gens voudraient pourtant habiter dans ...
NULL ({ 30 }) most ({ 1 2 3 4 }) people ({ 5 6 }) , ({ }) however ({ 8 }) , ({ })
would ({ 7 }) like ({ }) to ({ }) live ({ 9 }) in ({ 10 }) ...
```

Figure 1: French/English word alignment

is much higher than af(*demander, whether*). This needs to be taken into account in weighting.

Alignment frequency should not be the only measurement of significance. Since frequent bad translations (e.g., *the*) have a much higher af value than infrequent good translations (e.g., *question*) we also need to take corpus frequency cf into account.

We use the following weight to combine alignment frequency and corpus frequency:

$$\mathrm{weight}(t \mid v_i) = \max\left(\frac{\mathrm{af}}{n}, \frac{\mathrm{af}}{\mathrm{cf}}\right) \quad (1)$$

We then use cosine similarity to compute the similarity of vectors.

$$\mathrm{sim}(a, b) = \frac{a \cdot b}{||a||||b||} \quad (2)$$

This will yield a value $x \in [0, 1]$ where 1 means that both vectors are identical and 0 means that they share no similar terms.

3.2. Clustering

To partition the set of occurrences of a verb, we apply two variants of the K-means algorithm. In standard K-means, the parameter K, the number of clusters, is given. In our application, it is more important to avoid very small and very large clusters. Large clusters may "hide" rare and novel senses that would be included in a small cluster if the large cluster were to be subdivided further. Too many small clusters are time consuming to sift through and would be detrimental to our goal of making the identification of rare and novel senses quick and efficient.

To avoid both small and large clusters, we initially cluster the vector space into subclusters of a desired maximum size m and then group similar subclusters into superclusters. The parameter m ($8 \leq m \leq 14$ in this paper) allows to control the granularity of the clustering process.

We use *n-secting K-means*, a variant of bisecting K-means, to compute subclusters. N-secting K-means divides a subcluster ω into n new subclusters where $n = \max(2, |\omega|/m)$. We define $\omega = V$ for the initial state. The algorithm stops if all subclusters have at most m members. For the small values of m we use in this paper, this clustering yields a set of 10^3 – 10^4 subclusters for the high-frequency French verbs that our algorithm is intended for.

Finally, we employ standard K-means on the centroids to group subclusters into superclusters that users can easily navigate. We set $10 \leq K \leq 25$ based on the usability constraints of the user interface shown in Figure 3.

4. System Design

We chose a centralized web based approach. This means users can access the server from any browser. On the server side we employ Python (using Apache and mod_python) for the backend and the user interface.

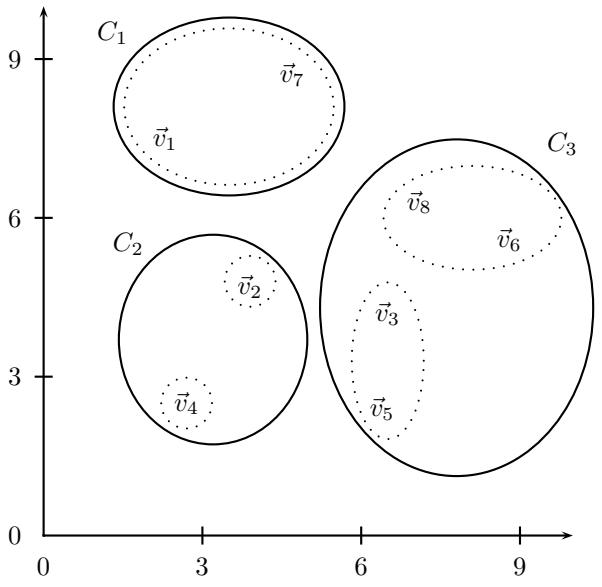

Figure 2: Example of subclusters and superclusters.

All calculated data is stored in XML format and can be downloaded by the user any time, e.g., for offline use with an external program.

Figure 3 shows how a supercluster is visualized. On top there are several boxes that each represent one of the final superclusters. The background color of each box gives an indication of the homogeneity of the corresponding cluster. The greener a box, the better the homogeneity. The homogeneity of a supercluster is the average similarity of its centroid to the centroids of its subclusters. Hovering over a box displays more information about the supercluster (26 subclusters etc. in the example). Underneath the boxes, the highest-weighted terms in the centroid are displayed (e.g., 0.470 for *abandonar*). This helps the user to understand what type of senses a supercluster contains.

Figure 4 shows the representation of a subcluster. The interface shows its size ("Members"), its homogeneity ("Integrity"), its similarity to the supercluster, and highly weighted terms in its centroid (box "Terms in the center ..."). Below that, a complete list of its member sentences is given. In each sentence, the word that GIZA++ determined to be a translation of the query is highlighted.

5. Evaluation

5.1. Evaluation against Dictionary Senses

An exhaustive evaluation is time-consuming as each occurrence of the verb in Europarl must be manually assigned to one of the senses. We performed this evaluation for

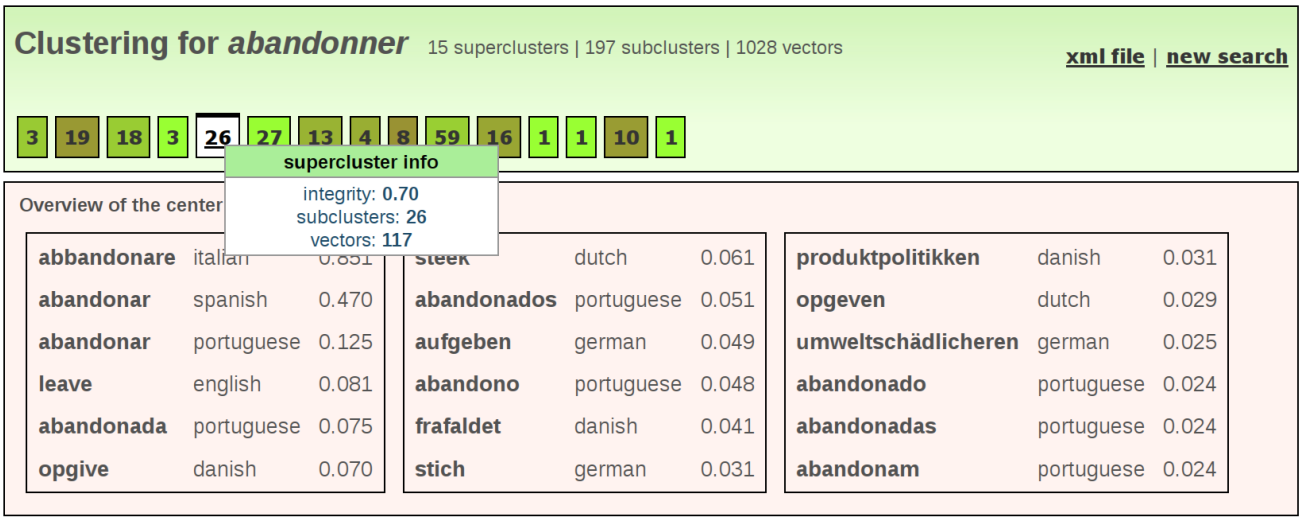

Figure 3: Overview of a specific supercluster

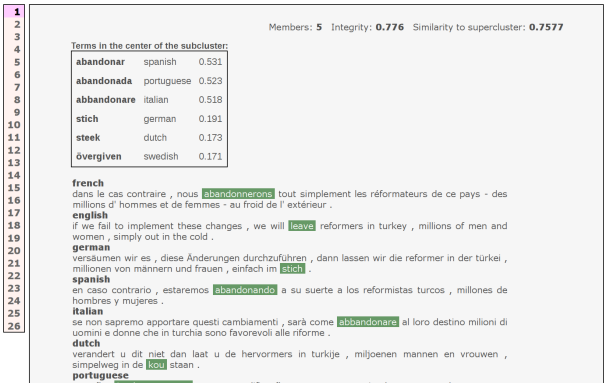

Figure 4: View of a subcluster

mesurer. 15 of 25 superclusters were dominated by one sense with one or two occurrences of another sense. Two superclusters were heterogeneous. One supercluster consisted of nominal uses of the verb root only – this is due to erroneous lemmatization, which we plan to fix in the future. Seven superclusters were the type of noise superclusters that we are looking for to identify rare and novel senses.

We interpret these results as an indication that the system provides a clustering with the desired properties: the bulk of the instances of the verb can be discarded because they occur in mostly homogeneous superclusters in which instances do not have to be inspected. The analyst can concentrate on heterogeneous superclusters to find rare and novel senses.

5.2. Qualitative Evaluation

We performed a qualitative evaluation for 6 verbs: *abandonner*, *glisser*, *mobiliser*, *parcourir*, *payer*, and *remercier*. For each verb, the evaluating linguist went through the superclusters one by one and discarded "green" or homogeneous ones. In each case, the color coding was confirmed by randomly checking a number of instances. The analysis then concentrated on noise superclusters. These are usually small, consisting of a number of instances that are different from all other instances in the corpus. The following rare or novel usages of the verbs were found:

- *abandonner qc en faveur de qc*
- *abandonner qc pour qc*
- *glisser qc à qn* (in the sense "give")
- *faire payer* (in the sense "charge")
- *payer qc sur qc* (in the sense "pay something on something else", 'la finlande oblige les personnes qui reviennent à payer sur leur vehicule des taxes d'importation')

These usages are not listed in the Petit Robert (2008), one of the standard dictionaries of French with fine sense distinctions and more than 300,000 senses. However, it is normal for dictionaries not to account for systematic causative constructions like *faire payer*, which may be considered as a typical case of lexical gap in French, exhibited only by translation equivalents like *charge* or *berechnen*.

5.3. Evaluation using pseudowords

Figure 5 shows an evaluation using pseudowords. All occurrences of the verbs *payer* and *abandonner* were replaced with the artificial word *a-p*. All instances of *a-p* were then clustered into superclusters. Each supercluster was assigned to the verb that was responsible for the majority of its members. Purity for a verb was defined as the average purity of the superclusters that the instances of the verb occur in. E.g., if two occurrences occur in a supercluster with purity 0.9 and one occurs in a supercluster with purity 0.6, then overall purity is 0.8.

The figure demonstrates that superclusters are more than 90% pure for $K \geq 10$ superclusters. For the intended application we always use 10 or more superclusters. This demonstrates that, if verb senses are sufficiently distinct, frequent standard senses will be well separated into pure superclusters, so that the lexicographer can concentrate on small noise superclusters to find rare and novel senses.

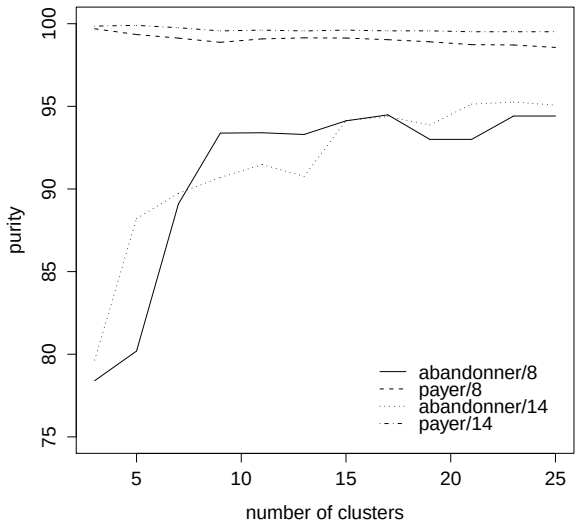

Figure 5: Evaluation using pseudowords

6. Conclusions and Future Work

We have presented a system for clustering multitext instances of French verbs. The system allows a lexicographer to discard most occurrences and quickly find rare and novel senses. We have shown that very distinct senses are successfully discriminated and that frequent and regular uses of the word are grouped into homogeneous superclusters that can be quickly discarded by the lexicographer. Several rare and novel usages of French verbs were found using the system. The method is not limited to just French verbs, but could be applied to different syntactic categories in any language.

The quality of the results depends heavily on the quality of the word alignment and the translation. Since in this project we deal with verbs and particularly long sentences, alignment results are not ideal.

Of course, rare and novel senses in the source language need not to be translated in a way that makes them distinguishable. Some target languages may not reflect the special use. This may result in rare uses being hidden in clusters otherwise populated by frequent uses. Also if a particular use is very rare, the word alignment tends to produce incoherent results that do not necessarily allow fair conclusions. It also should be noted that the domain restriction of the corpus potentially limits the variety of senses that can appear.

On the other hand it is also possible for frequent and regular uses to be assigned to inhomogeneous clusters – if the alignment or translation so permits. Additionally the clustering process itself is designed to be fast and effective, but does not guarantee to find the best possible cluster configuration.

A possibility to lower the severity of these effects is generally to use more languages. Manually fine tuning the parameter settings for the clustering process of a query can also yield better results.

These problems are part of the reason why it is difficult to expand the system into a fully automated system that does not depend on a lexicographer. Besides, even for human judgement to decide whether a particular use is regular or rare is not trivial. A statistical approach struggles even more when the distinction is not clear-cut.

For a fully automated system to be feasible, the word alignment would have to be very good even for infrequent words and work would have to be put into further enhancing the clustering process.

7. References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. A statistical approach to sense disambiguation in machine translation. In *DARPA Workshop on Speech and Natural Language*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *ACL*.

Jacques Francois. 2007. *Pour une cartographie de la polysemie verbale*. Peeters, Leuven, Paris.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Nancy Ide, Tomaz Erjavec, and Dan Tufiş. 2002. Sense discrimination with parallel corpora. In *ACL workshop on Word sense disambiguation*.

Nancy Ide. 1999. Parallel translations as sense discriminators. In *ACL SIGLEX*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit*, pages 79–86.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In *ACL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *EACL*.

2008. *Le Nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française*, volume Nouvelle édition remaniée et amplifiée sous la rédaction de Josette Rey-Debove et Alain Rey. Le Robert, Paris.

Sabine Ploux and Bernard Victorri. 1998. Construction d'espaces semantiques a l'aide de dictionnaires de synonymes. *Traitement automatique des langues*, 39(1):161–182.

Hinrich Schütze. 1993. Translation by confusion. In *AAAI Spring Symposium on Building Lexicons for Machine Translation*.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *COLING*.

Jakob Uszkoreit and Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *ACL/HLT*.