Discovering semantic relations by means of unsupervised sense clustering

Marianna Apidianaki

Language and Translation Technology Team (LT³)
Faculty of Translation Studies, University College Ghent, Belgium
marianna.apidianaki@hogent.be

Abstract

Electronic sense inventories are needed for Word Sense Disambiguation in Natural Language Processing (NLP) applications. However, existing sense inventories are criticized as being inadequate for this task. The word senses proposed in these resources are listed without any information on their distinguishability and their relations. Furthermore, the fine granularity of the senses is not always needed in NLP applications. Nevertheless, this type of knowledge can be automatically extracted from textual data using machine learning techniques.

In this paper, we show how an unsupervised sense induction method permits to capture two different types of semantic relations: first, the semantic similarity relations between the translation equivalents of polysemous words; second, the relations between their senses, which are automatically identified by means of semantic clustering. We analyze the results of this method and compare them with the semantic descriptions provided in a typical multilingual semantic resource.

1. Introduction

Pre-defined sense inventories are often criticized as being inadequate for Word Sense Disambiguation (WSD) in Natural Language Processing (NLP) applications. The main reason is that these resources contain a high number of too fine-grained senses, which are listed without any information on their relations. This lack is partly due to the incapacity of the sense enumeration techniques used in lexicography to justify a distinction between different types of ambiguity (Dolan, 1994; Pustejovsky, 1995). Additionally, it is explained by the fact that the majority of the resources were initially developed for use by humans, who can identify word sense relations even if the relevant information is not explicitly mentioned.

Nevertheless, the fine granularity of the semantic descriptions found in pre-defined resources does not seem to be necessary for efficient WSD. According to Ide and Wilks (2007), NLP applications, when they need WSD, seem to need homograph-level disambiguation. Finer-grained distinctions are rarely needed, and when they are, more robust and different kinds of processing are required. Furthermore, the fine granularity of the semantic descriptions found in existing resources, combined to the great divergences in their structure and content, put a hindrance to their compatibility. Their exploitation in multilingual applications is rather problematic as well, given the difficulty to establish correspondences between fine-grained senses of words and their translation equivalents in other languages (Specia et al., 2006).

However, the semantic information needed for WSD can be acquired directly from texts without recourse to pre-defined inventories. In this paper, we first analyze the problems raised during WSD by the fine granularity of word senses and we explain how it is possible to obtain coarser senses from pre-defined inventories. Then, we argue in favor of data-driven semantic analysis methods. We show how an unsupervised sense induction method can reveal two types of semantic relations: a) the relations between the translation equivalents (TEs) of polysemous words of one language in another language, and b) the relations between the senses of the polysemous words, which correspond to sense-clusters of their TEs. The results of this method are analyzed and compared to the descriptions provided in a predefined multilingual semantic resource.

2. Using pre-defined resources for WSD

2.1. Sense granularity

Electronic sense inventories provide the list of the candidate senses of polysemous words that are needed for WSD. The task of a WSD algorithm is to select, from this list, the most appropriate sense for each new occurrence of a polysemous word in texts. However, this selection is complicated when the WSD algorithm is confronted to a high number of fine-grained senses. This criticism has been mainly formulated by reference to WordNet, a resource widely exploited in WSD tasks (Edmonds and Kilgarriff, 2002; Ide and Wilks, 2007).

The pertinence of the selection of one among close senses can be doubted as well, as it can lead to arbitrary decisions in cases where the occurrences of the polysemous words could correspond to more than one specific fine-grained sense. Given that inter-sense relations are not described, the selection would provoke a loss of useful information on the semantics of the word (Dolan, 1994).

The difficulty of selecting one among fine-grained senses for a new occurrence of a polysemous word is also observed in human annotation tasks. The inter-annotator agreement for word-sense tagging is lower when language users are asked to assign refined sense tags – such as those found in WordNet – especially when the definition entries are short and only a few or no example sentences for the usage of each word sense are provided (Veronis, 1998; Ng et al., 1999).

2.2. Sense clustering

The criticisms addressed to existing semantic resources, combined with some scepticism concerning the need for fine-grained semantic distinctions in NLP applications, enhanced the development of methods for deriving coarser sense inventories from existing ones (Peters et al., 1998;
Mihalcea and Moldovan, 2001; Navigli, 2006; Navigli et al., 2007). These sense clustering methods, developed by reference to WordNet and EuroWordNet, discover the relations between the sets of synonyms (synsets) that describe the senses of the words and collapse them into clusters. These methods perform clustering by exploiting different types of information: information concerning the similarity of the words found in the synsets describing different senses; information on the similarity of the relations between them and other synsets of the network; probabilistic information extracted from corpora; syntactic criteria concerning alternations with similar subcategorization frames; semantic criteria concerning the semantic class of arguments, the subject domain and the underlying predicate-argument structures (Resnik, 1995; Jiang and Conrath, 1997; Mihalcea and Moldovan, 2001; Palmer et al., 2006). The sense granularity reduction performed has been shown to improve the performance of WSD. It also facilitates the establishment of correspondences between the senses described in different resources, increasing their compatibility. Nonetheless, these knowledge-based clustering methods have been developed by reference to specific resources and cannot be generalized. This limitation and the increasing availability of text corpora have fostered the development of unsupervised methods capable of acquiring information on sense relations directly from texts.

3. Data-based semantic analysis

3.1. Discovering word sense relations from texts

Methods for discovering word sense relations from textual data have been developed in a monolingual as well as in a bi- and multi-lingual context. Monolingual methods are based on the distributional hypotheses of meaning and of semantic similarity, according to which semantically similar words present similar distributional behavior (Miller and Charles, 1991). The cooccurrences of the words in texts, or the syntactic frames in which they occur, constitute their sets of context features. The similarity of these sets reveals the similarity of the corresponding words (Church and Hanks, 1990; Dagan et al., 1993; Pereira et al., 1993; Pantel and Lin, 2002).

In a bi- (or multi-) lingual setting, word sense relations can be discovered by using translational information. In this case, the context of the SL words, that serves to calculate their similarity, corresponds to their translation equivalents (TEs) in other languages. The TEs are used to build vectors for the SL words whose similarity shows their semantic relatedness (van der Plas and Tiedemann, 2006). However, apart from revealing the relations of SL words, their TEs can also serve to analyze their semantics.

3.2. Cross-lingual sense induction

In the cross-lingual approach to sense induction, the TEs of the instances of polysemous words in parallel corpora can be used for identifying their senses. Ide et al. (2002) and Tufis et al. (2004) build vectors for the instances of SL words in a multilingual corpus by using as features their TEs in different languages. The vectors and the corresponding SL instances are clustered and the generated clusters describe the senses of the words. However, this method needs parallel corpora in many different languages, which are hard to be found. Additionally, the clustering is performed by an agglomerative algorithm which creates disjoint clusters. Consequently, the acquired senses are disjoint and their relations are not taken into account.

Another common approach to cross-lingual sense induction consists in using each TE of a polysemous word as describing one distinct sense. The most important merits of this approach are that it permits to bypass the subjectivity issue inherent in sense identification tasks and to derive senses relevant for translation (Resnik and Yarowsky, 1998). Nevertheless, the TEs of polysemous words may not always constitute valid sense indicators.

In cases of parallel (or translation) ambiguity, for instance, the TEs present the same ambiguity as the SL word. Ide and Wilks (2007) describe cases where the same historical processes of sense “chaining” occurs in different languages and the words extend their original sense in the same way. Relying on cross-language lexicalization for sense distinction in such cases would disregard the sense deviations characterizing the words and would lead to the conclusion that both have a single sense.

Another danger in using TEs as straightforward sense indicators is that they may carry senses valid only in the TL which should not indicate a sense split in the SL. This happens, for instance, when a generic word in one language describes senses expressed by distinct lexical elements in another.

Additionally, given that translators often use synonyms and near-synonyms in order to avoid repetitions in the translated texts, it is common that the TEs of a polysemous word be semantically similar in the target language (TL). Consequently, these TEs translate the same sense of the SL word and should not be used to induce senses in the SL.

Finally, the senses induced by using the TEs as straightforward sense indicators are uniform: clear-cut and finer sense distinctions are listed without any description of their relations. The theoretical and practical problems posed by this cross-lingual approach to sense identification are dis-

---

1 In EuroWordNet (Vossen, 1999), where the fine-grained English WordNet served as an Interlingual Index (II), the creation of coarse-grained inter-lingual entries facilitated the correspondences between equivalent senses in different languages.

2 It has been adopted in the multilingual tasks of the Senseval (Chklovski et al., 2004) and SemEval (Jin et al., 2007) exercises and in works on WSD in Machine Translation (Cabesaz and Resnik, 2005; Carpuat et al., 2006).

3 For example, the TE of the English noun interest in French, intérêt carries the “financial” and “personal” senses of the English word (Resnik and Yarowsky, 1998).

4 The English wing and its equivalent ali in Italian which both have extended their original sense from birds to airplanes, to buildings and to soccer positions.

5 For example, Japanese has different words for "wear", depending on what part of the body is involved, but this distinction is not performed in English (Gale et al., 1993).

6 According to Baker (1996), this particularity can be considered as a universal feature of translated texts.
also captures the relations between the senses of the polysemous words. An interesting property of the SEMCLU algorithm, which constitutes the core of this method, is that it permits to perform a fuzzy clustering. This means that the resulting sense clusters are not disjoint but may present overlaps. The overlaps of the clusters are formally described by the non-empty intersection of their elements. This property of the algorithm is very important for sense induction: given that the clusters describe word senses, their overlaps can be perceived as describing the relations between the corresponding senses.

Capturing word sense relations gives the possibility to perform a differentiation between close and distant (or antagonistic) senses. Furthermore, the proximity of the senses can serve to modify the granularity of the proposed sense descriptions. Overlapping clusters often describe nuances or sub-senses of coarse-grained senses. Consequently, their fusion makes it possible to obtain a description of the main sense distinctions characterizing a word.

A side benefit of this type of representation is that it allows to take into account the cases of translation ambiguity. Such cases are observed when a TE is found in the intersection of clusters. This means that the TE is polysemous, that it carries the senses described by the clusters, which may be distant or close, and that it translates both in the TL.

4.3. Clustering examples

The sense induction method presented above builds bilingual sense-cluster resources for different language pairs automatically. What is needed is a parallel training corpus and tools for sentence and word alignment, part-of-speech tagging and lemmatization, which are available in many languages.

In this paper, we focus on the clustering performed for two polysemous English nouns (movement and plant). Our intention is to analyze the semantic representations obtained by this method from a qualitative point of view. The results presented here are obtained by training the sense induction method on the English–Greek part of the multilingual INTERA parallel corpus (Gavrilidou et al., 2004).

- movement

The English noun movement has six TEs in Greek in the training corpus: κυκλοφορία (kykloforia / 251), διακίνηση (diakinisi / 38), κίνηση (kinisi / 28), μετακίνηση (metakinisi / 19), κίνημα (kinima / 11), κινητικότητα (kinikitotita / 6).\(^7\) In the "traditional" approach to cross-lingual sense induction, each TE would describe one distinct sense of the English word and, consequently, movement would be considered as having six distinct senses. However, this analysis would not be correct because some of the TEs of the word are semantically related in the TL.

---

\(^7\) The TEs that translate less than five instances of the SL word in the training corpus are not considered, as the available contextual information is not sufficient for efficient processing. In parentheses, we give the transliteration of the Greek TEs and the number of instances of the SL word that they translate in the corpus.
In the clustering solution generated by SEMCLU, the semantically similar TEs of *movement* are grouped into clusters which describe its senses. The clustered TEs translate occurrences of *movement* that are semantically related, i.e. found in similar contexts. The clusters obtained for *movement* are described in Table 1 and schematically illustrated in Figure 1.

We observe that the clusters (a), (b) and (c) share some elements and, consequently, they overlap. The TEs found in the overlapping clusters translate the "physical movement" sense of the word and each separate cluster describes a nuance of this coarse sense. However, the cluster (d) is clearly distinguished from the others. Its unique element (*κίνημα*) translates the sense of "social movement".

The "physical" and "social" senses of *movement* are more distant than those described by the overlapping clusters. This is clearly illustrated in Figure 1. The relations and the distinctions identified between the clusters reflect the different status of the corresponding senses. In the following section, we will show how this information can be exploited in order to modify the granularity of the obtained senses.

### Table 1: Sense-clusters of "movement" and "plant"

<table>
<thead>
<tr>
<th>polysemous word</th>
<th>sense-clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>movement</td>
<td>a. {μετακίνηση (metakinisi)}, κίνηση (kinisi), διακίνηση (diakinisi)}</td>
</tr>
<tr>
<td></td>
<td>b. {κίνηση (kinisi), διακίνηση (diakinisi), κυκλοφορία (kykloforia)}</td>
</tr>
<tr>
<td></td>
<td>c. {μετακίνηση (metakinisi), διακίνηση (diakinisi), κινητικότητα (kinitikotita)}</td>
</tr>
<tr>
<td></td>
<td>d. {κίνημα (kinima)}</td>
</tr>
<tr>
<td>plant</td>
<td>a. {μονάδα (monada), εγκατάσταση (egkatastasi)}</td>
</tr>
<tr>
<td></td>
<td>b. {σταθμός (stathmos), εργοστάσιο (ergostasio)}</td>
</tr>
<tr>
<td></td>
<td>c. {σταθμός (monada), μονάδα (monada)}</td>
</tr>
<tr>
<td></td>
<td>d. {φυτό (fyto)}</td>
</tr>
</tbody>
</table>

---

**Figure 1: Sense-clusters of "movement"**

The TEs of *plant* in the training corpus are the following: *φυτό* (fyto / 94), *μονάδα* (monada / 15), *εγκατάσταση* (egkatastasi / 14), *εργοστάσιο* (ergostasio / 14), *σταθμός* (stathmos / 7). The clusters obtained for this word are described in Table 1 and illustrated in Figure 2.

As in the case of *movement*, some of the obtained sense-clusters overlap while one is disjoint. The relations and distinctions of the clusters indicate the relations between the corresponding senses of the SL word. *Plant* constitutes a case of homonymy where the implicated lexical units carry two distinct senses, the "botanical" and the "industrial" senses. The "industrial" sense is translated by the TEs found in the overlapping clusters ((a), (b) and (c)), which describe nuances of this coarse sense, while the "botanical" sense is described by the disjoint cluster (d). The TE found in this cluster (φυτό), is the most frequent TE of *plant* in the corpus and the only that translates its "botanical" sense. So, the distinction between the disjoint and the overlapping clusters reflects the coarse distinction between the two non-related senses of *plant*.

It is important to note that the isolation of a TE in a disjoint cluster does not always indicate a clear-cut semantic distinction. This isolation may also be provoked by the low-frequency of a TE in the corpus, because of data sparseness. Nevertheless, the high frequency of a TE (like φυτό) constitutes a reliable clue for spotting a semantic distinction, given that the relative contextual information is sufficient for identifying the semantic relation of the TE to the other TEs of the SL word, if such a relation exists.

The creation of overlapping and disjoint clusters reflects the differences in the nature and the status of the obtained senses. This information on the distinguishability of the senses and their relations is important from a lexicographic point of view, as it offers the possibility to identify different types of ambiguity that are not taken into account in other semantic resources. It also permits to obtain a dynamic representation of the semantics of words that can be adapted to the uses of the inventory. In the next section, we show how the granularity of the senses can be modified by exploiting the information on their relations found in the sense-clustering results.

### 4.4. Sense granularity modification

The semantic representations generated by the sense induction method presented in this paper give the possibilit-

---

8As the contextual information used concerns first-order cooccurrences, the method is rather sensible to data sparseness. The use of more abstract information concerning higher-order occurrences constitutes an avenue for future work.
ity to automatically modify the granularity of the obtained senses. The overlapping clusters describe fine-grained senses of the SL words or nuances of coarse senses. The clusters may also overlap because of the lack of pertinent links between some of their elements, which may be due to data-sparse ness and not to the absence of a semantic relation. These clusters can be merged on the basis of their overlaps into bigger ones, describing coarser-grained senses.

In the case of movement, the merging of the overlapping clusters ((a), (b) and (c)) would generate a bigger cluster describing the "physical movement" sense. The distinction between this cluster and the cluster (d), which describes the "social movement" sense, reflects the main semantic distinction characterizing the SL word.

1. movement - \{{{\text{μετακίνηση}}, \text{κίνηση}, \text{διακίνηση}}, \{{{\text{κίνηση}}, \text{διακίνηση}}, \text{κυκλοφορία}}\}

2. movement - \{{{\text{κίνημα}}\}

In the case of plant, the fusion of the overlapping clusters ((a), (b) and (c)) generates a bigger one describing the "industrial" sense of the word, which is contrasted to the "botanical" sense, described by (d). These two clusters reflect the homonymous distinction characterizing plant.

1. plant - \{{{\text{μονάδα}}, \text{σταθμός}}, \{{{\text{μονάδα}}, \text{εγκατάσταση}}, \{{{\text{σταθμός}}, \text{εργοστάσιο}}\}

2. plant - \{{{\text{φυτό}}\}

4.5. Discussion

We have shown how the inter-sense relations captured by the clusters overlaps can serve to modify the granularity of the obtained senses. Instead of going "bottom-up" (i.e. grouping fine-grained senses to form coarser ones), it would also be possible to adopt a "top-down" approach (i.e. to move from coarse-grained senses, described by disjoint clusters, to fine-grained ones). As there is no unique answer to the question of the granularity of the word sense descriptions needed in different NLP tasks, the possibility to modify it automatically permits to adapt the descriptions to the needs of specific applications. Taking into account the proximity of the senses facilitates the task of WSD algorithms as well, which do not have to make a selection among a large number of close but distinct, and hardly distinguishable, senses.

Additional benefits become evident during WSD evaluation. When the relations between senses are not taken into account, errors concerning close or distant senses are considered as equally important. Considering the inter-sense relations makes possible a differing penalization of WSD errors (Resnik and Yarowsky, 1998) and renders the evaluation more flexible and sophisticated. The advantages of exploiting this type of semantic representation in multilingual WSD, and in WSD and Machine Translation evaluation, are presented in Apidianaki (2009) and Apidianaki et al. (2009).

5. Evaluation

5.1. Difficulties

The evaluation of automatic sense induction methods is difficult due to the lack of a gold standard in lexical semantics and the great divergences in the content and the coverage of existing hand-crafted semantic resources. Another important issue is that WSD constitutes an intermediate task in NLP applications, which aims to improve their performance (Wilks and Stevenson, 1996). Consequently, different applications have varying WSD needs which have an impact on the semantic descriptions that should be used. So, the results of an evaluation of the contents of a semantic inventory would not always be meaningful for the usefulness of the inventory in different settings.

An extrinsic evaluation of the results of the sense induction method presented in this paper has been performed in Apidianaki (2009). In this work, it is shown how the exploitation of the automatically built sense-cluster inventory can improve the performance of a WSD method in a bilingual context. Furthermore, the work of Apidianaki et al. (2009) shows how the use of this type of inventory can be beneficial in MT evaluation. Here, we present a more focused qualitative evaluation of the clustering results, by comparing them to the contents of an existing multilingual semantic resource.

5.2. Qualitative evaluation

We compare the senses obtained by the clustering method to the ones found for the same polysemous English nouns in BalkaNet (Stamou et al., 2002; Tufis et al., 2004). BalkaNet is a multilingual semantic network that comprises wordnets in six different languages (Greek, Bulgarian, Romanian, Turkish, Serbian and Czech). Each wordnet contains concepts organized in semantic taxonomies, which are put into correspondence with their semantic equivalents in the other languages via an Interlingual Index (ILI) composed by concepts of Princeton WordNet. The ILI connects the languages between them and makes possible the transition from the concepts of one language to semantically similar concepts in another.

The senses found in the BalkaNet ILI for movement and plant are described in Table 2. We present the ILI synsets describing each sense of the words, the corresponding Greek synsets and the provided sense descriptions (glosses), which are common for the synsets of the two languages.

- movement

The first three senses given for movement in BalkaNet are too fine-grained and hardly distinguishable. The words found in the ILI synsets are very similar and the corresponding glosses as well. It is clear that the three synsets refer to the "physical movement" sense but their differences are minimal. The proposed semantic distinctions cannot be clarified neither by looking at the glosses, which are very similar too, nor by using the Greek synsets corresponding to the ILI synsets.

The Greek synsets corresponding to the first two senses contain the same equivalent, meaning "change of posi-
So, we observe that the senses described by the synsets are too fine-grained and related, and that they could be grouped in one, describing the coarser sense of "physical movement". The clustering of the synsets could be done by exploiting their relations which could be identified by methods such as the ones described in section 2.2. The semantic relations of these synsets could be discovered on the basis of their similarity and relatedness. The synset (1) and (2) are considered as performing the same function as the others, is situated at the same level as the three related senses of movement. Like WordNet, BalkaNet is based on the enumeration approach, which does not operate distinctions between close and distant senses and different types of ambiguity. So, the distinct senses of an entry are not organized hierarchically but listed sequentially. The distinct senses of the synset could be performed by exploiting taxonomic information (like the hyperonym of synset 4, which is the synset "social group"). However, even if it is possible to discover the relations and distinctions of the senses on the basis of information available in the network, these are not explicitly described in the inventory. The senses are organized by frequency of use and there is no indication of the degree of distinguishability between them. These lacks result in the uniform processing of the proposed senses during WSD and have a negative impact during WSD evaluation, as errors concerning close and distant senses are equally penalized (Resnik and Yarowsky, 1997; Resnik and Yarowsky, 1998). The relations and distinctions between the senses are explicitly described in the clustering solution obtained for movement. The coarse-grained senses acquired after merging the overlapping clusters correspond to the senses that would be obtained if the three fine-grained synsets of the word were grouped. The merging of the clusters could, thus, be considered as performing the same function as the knowledge-based methods that have been proposed for reducing the granularity of WordNet senses.

### plant

The first two synsets of plant in BalkaNet correspond to its "botanical" and "industrial" senses. As we have shown, this coarse distinction between the homonymic senses of the word was identified in the sense-clustering after the modification of the granularity of the senses.

---

The proposed TE is not found in the clustering results because only one-word TEs were retained.

---

9 The proposed TE is not found in the clustering results because only one-word TEs were retained.
The senses described by the synsets (3) and (4) are too rare. Instances of plant carrying these senses do not appear in our training corpus, so no information about them is found in the sense-clusters. The suggestion of too rare senses, or of senses not relevant to the domains of the processed texts, constitutes a drawback of exploiting predefined resources for WSD (Pantel and Lin, 2002). Considering such senses increases the number of possible choices during WSD which makes processing more complex without any benefit.

The synsets (3) and (4) that describe these rare senses contain no synonyms and are not related to other synsets in the network. Furthermore, it would be difficult to consider the TEs found in the corresponding Greek synsets (παγόδα, (“trap”) and τραπεζαίος (“actor”)) as translations of plant in Greek.

The varying quantity of information provided for the different synsets of the word is justified by the different status of the concerned senses. Nevertheless, even the information found in the Greek synsets corresponding to the main senses of plant, the first two, is limited. Each synset contains only one TE: the synset φυτό describes the "botanical" sense of plant and the synset μονάδα, εγκατάσταση, μονάδα, εγκατάστασεις (meaning "industrial installations") describes its "industrial" sense. In this case too, the clusters describing these senses contain richer information than the BalkaNet synsets. The cluster of the TE εγκατάσταση contains the semantically similar word μονάδα, and if the overlapping clusters are merged, more related words are included: the TEs σταθμός and εγκαταστάσεις. Additionally, contrary to BalkaNet, where the senses are situated at the same level and treated as uniform, in the sense-clustering results the senses are differentiated: the mutually exclusive senses are described by distinct clusters while the related ones are described by overlapping clusters.

6. Conclusion

In this paper, we have shown how word sense relations can be discovered from textual data by means of unsupervised learning. The sense induction method used identifies the semantic relations between the translation equivalents of polysemous words in a parallel corpus and groups them into clusters, which describe the senses of the source language words. Using the clusters of equivalents as sense indicators increases the pertinence of the identified senses in comparison to the traditional cross-lingual approach to sense induction, where each equivalent indicates a distinct SL sense.

In the generated semantic inventory, the senses of polysemous words are described by clusters of their translation equivalents in another language. A merit of this approach is that the acquired senses are not simply listed, as in pre-defined semantic resources, but their relations are taken into account as well. The description of the relations between word-senses permits to automatically modify their granularity and to adapt them to the WSD needs of specific applications. Further advantages of this unsupervised semantic analysis method are its language-independency and the relevance of the acquired descriptions to the processed data. Directions for future work include the automatic elaboration of sense-cluster inventories for different language pairs and their exploitation for WSD in multilingual applications.

7. References


