Abstract

In this paper the phase of semantic valence dictionary of Polish verbs consisting in aggregating entries to semantically coherent sets is presented. Two methods: a simple agglomerative one and minimal spanning trees method are discussed and compared. Both methods use a predefined similarity measure of semantic frames.

1 Introduction

The primary task of our research is to create a semantic valence dictionary in an automatic way. To accomplish this goal, the valence dictionary of Polish verbs is supplemented with semantic information, provided by wordnet’s semantic categories (Hajnicz, 2009d; Hajnicz, 2009c) or synsets (Hajnicz, 2009a) of nouns. In our present work we focus on arguments taking form of nominal phrases NPs and prepositional-nominative phrases PrepNPs, whose semantic heads are nouns. We discuss the case of 26 predefined semantic categories of nouns, which is simpler than the case of actual wordnet synsets. In the current phase of work we want to discuss in this paper, we have in our disposal two resources:

- purely syntactic valence dictionary,
- a syntactically and semantically annotated corpus.

In theory, it is not important whether these resources were prepared manually or automatically. In practice, the difference is quite significant, because errors obtained during automated data processing are cumulated.

Typical approaches, e.g., VerbNet (Dang et al., 1998) or VerbaLex (Hlaváčková and Horák, 2006), consider one strongly preferred sense per argument. In contrast, we present a solution in which all appropriate senses are aggregated.

2 Data resources

We used an extensive valence dictionary based on Świdziński’s (1994) valence dictionary containing 1064 verbs. It was specially modified for our task. Świdziński’s dictionary was supplemented with 1000 verb entries from the dictionary automatically obtained by Dębowski and Woliński (2007) to increase the coverage of used dictionary on SEMKIPI (cf. below). The most carefully elaborated part of the valence dictionary concerns the set of 32 verbs manually chosen for the experiments (Hajnicz, 2009c). They were chosen manually in order to maximise the variability of their syntactic frames (in particular, diathesis alternations) on one hand and the polisemy within a single frame on the other. Their frequency was the important criterion for this choice as well.

A syntactic dictionary $D$ is a set of entries representing schemata for every verb considered. Formally, $D$ is a set of pairs $\langle v, g \rangle$, where $v \in V$ is a verb and $g \in G$ is its syntactic schema. Below we list syntactic dictionary entries for verb interesować (to interest). np:case are nominal phrases, sentp:wh are wh-clauses, whereas sie is a reflexive marker.

(1) interesować np:acc np:nom
   interesować np:inst np:nom sie
   interesować np:nom sentp:wh sie

The main resource used in our experiments was the IPI PAN Corpus of Polish written texts (Przepiórkowski, 2004). A small subcorpus was selected from it, referred to as SEMKIPI containing 195 042 sentences predicated by chosen verbs. SEMKIPI was parsed with the Świgra parser (Woliński, 2004) based on the metamorphosis grammar GFJP (Świdziński, 1992) provided with the valence dictionary presented above.¹ The complete frequency list of verbs in the IPI PAN Corpus contains about 15 000 verbs, with 12 000 of them occurring at least 5 times. Grammatical dictionary of Polish (Saloni et al., 2007) lists 29 000 verbs.

In order to reduce data sparseness, in the present experiment we considered only the topmost phrases being the actual arguments of a verb (i.e., a subject and complements included in its valence schema). This means that each obtained

¹In particular, the parser links genitive of negation with accusative in the corresponding valence schema.
The process of collecting a semantic valence protodictionary on the basis of SEMKIP! for semantic categories was described in (Hajnicz, 2009b).

Formally, a semantic protodictionary $D$ is a set of tuples $\langle (v,g,f), n_g, m_f \rangle$, where $(v,g) \in D$ is a schema of a verb, $f \in F_1$ is one of its semantic frames, $n_g$ is the frequency of $(v,g)$ and $m_f$ is the frequency of $(v,g,f)$. A frame is a list of arguments, among which only NPs and PrepNPs are semantically interpreted, i.e., supplied with semantic categories $c \in C$.

An exemplary subset of the set of frames connected with the schema np:acc np:dat np:nom of the verb proponować (to propose) is shown in (2). In the second column the frequencies of frames are given.

(2) proponować
np:acc: act; np:dat: person; np:nom: person 573
np:acc: act; np:dat: person; np:nom: person 51
np:acc: act; np:dat: group; np:nom: person 50
np:acc: act; np:dat: act; np:nom: person 31
np:acc: act; np:dat: person; np:nom: group 22
np:acc: act; np:dat: group; np:nom: person 16
np:acc: act; np:dat: location; np:nom: person 9
np:acc: act; np:dat: act; np:nom: group 8
np:acc: act; np:dat: feeling np:nom: group 4
np:acc: act; np:dat: group; np:nom: event 1

4 The process of aggregation
A protodictionary has plenty of entries (simple semantic frames), with a single category assigned to each syntactic slot. This does not reflect the actual semantics of a verb, since different categories of arguments do not entail different meanings of the verb. In other words, such classification is too fine-grained. For instance in sentences (3) we have different meanings of the verb przejechać. These differences are reflected in different English translations of the verb: to cross in the first sentence and to run over in the second. Hence, we want to have two different entries for it in the valence dictionary, with location and animal on the object position, correspondingly. On the other hand, in sentences (4) we deal with the same meaning of the verb kupić (to buy), and we want to have one entry for it. In order to differentiate these situations we defined a similarity measure $d$ between two categories. Its value varies from $1$ to $6$ for two “neighbouring” categories. The similarity measure between semantic categories is presented in Figure 1 in a form of graph in which nodes represent categories. $d(c_1, c_2)$ is the shortest path linking categories $c_1$ and $c_2$, interpreted as a sum of edges labels.²

Usage of the measure is based on the assumption that two categories are put together only if all categories located in between by means of a particular similarity measure occur at a considered slot of a schema as well. Observe that one can buy almost everything, in particular things having semantic categories positioned in between animal and location (in particular, food, substances, artifacts, some physical objects and groups of things, cf. Figure 1). Contrary, objects of crossing and running over are separated.

Synsets for which there is not a path in hyponymy relation and that are not top ones are not similar by definition.

(3) Piotr przejechał park samochodem (Piotr cross his park in a car)
Piotr przejechał psa samochodem (Piotr ran over his dog by a car)

(4) Piotr kupił bratu park (Piotr bought his brother a park)
Piotr kupił bratu psa (Piotr bought his brother a dog)

Thus, we want to aggregate simple frames into compound ones, in which every syntactic slot is supplied with a list of semantic categories. A compound frame is supposed to determine a single meaning of a verb. To obtain this, we have applied two clustering methods. Both are based on a similarity measure between frames $D_n$, where $n$ is a space dimension (number of NPs/PrepNPs). $D_n$ is defined on the basis of similarity measure between categories $d$ applied for all NPs/PrepNPs in Euclidean way. Namely,

$$D_n(f^A, f^B) = \sqrt{\sum_{i=1}^{n} (d(c^A_i, c^B_i))^2}$$

for $g = \langle r_1, \ldots, r_n \rangle$ and $f^A = \langle \langle r_1, c^A_1 \rangle, \ldots, \langle r_n, c^A_n \rangle \rangle$, $f^B = \langle \langle r_1, c^B_1 \rangle, \ldots, \langle r_n, c^B_n \rangle \rangle$.

The first method is a simple agglomerative method (Aggl) based on choosing the most frequent simple frame and joining it with other elements of a compound frame under creation that

²Please note that the graphical composition of a picture is not meaningful; in particular, the length of arcs is not proportional to the actual distance between nodes. Observe that the measures are not 2D; there are only visualised on a plane.
Figure 1: Similarity measure between semantic categories

(5) proponować np:acc np:dat np:nom
acc: act, event, place, state, time;
dat: cognit., communic., feel., group, person, poss., quality, relation;
nom: group, person, relation
264
acc: act, place, state;
dat: act, event, place, state, time;
nom: group, person
105
acc: cognit., communic., feel., group, person;
dat: group, person;
nom: group, person, relation
49
acc: act;
dat: artifact;
nom: group, person
22
acc: act;
dat: group, person;
nom: artifact
8
acc: act, event;
dat: group;
nom: act, event
7
acc: act;
dat: quantity;
nom: group
5
acc: act;
dat: act;
nom: artifact
4
acc: act;
dat: cognit.;
nom: artifact
2
acc: act;
dat: quality;
nom: artifact
2
acc: act;
dat: quantity;
nom: quantity
1

are “sufficiently” similar, i.e., $D_n$ does not exceed a particular threshold $\rho^A$.

A fragment of the aggregated dictionary $\tilde{D}$ for the schema np:acc np:dat np:nom of the verb proponować (to propose) is shown in (5).

The second method is a popular clustering method based on similarity measure called minimal spanning trees (MST) proposed by Zahn (1971). The algorithm was performed for each verb schema independently. Simple frames represented graph nodes, and edges were labelled with distances defined by $D_n$. The heuristics for determining threshold used for removing outlying edges $\rho^q$ was based on local criteria (the median $\mu_{(v,g)}$ and $q$’s percentile $\Phi_{(v,g)}^q$ of a distribution of lengths of edges between frames of a particular syntactic schema) and global criteria (the median $\mu_n$ and $q$’s percentile $\Phi_{n}^q$ of a distribution of lengths of edges between frames of all syntactic schemata with $n$ NPs/PrepNPs). Namely,

$$\rho_{(v,g)}^q = \max(\mu_n, \mu_{(v,g)}), \min(\Phi_n^q, \Phi_{(v,g)}^q)).$$

Medians ensure that too short edges will not be cut, percentiles ensure that too long edges will not stay.

5 Experiments

The experiments were performed with $\rho^A = 2$ for agglomerative method and percentiles $q = 80, 90$ for MST. Observe that the greater $\rho^A$ (or the higher $q$) the larger compound frames are obtained.

5.1 Manually prepared semantic dictionary

$D^H$ differs from $\tilde{D}$ in that it has no frequencies assigned to frames. Moreover, it is rather exhaustive, i.e., frames contain all corresponding semantic categories of slots. This means that such a dictionary should be interpret in a manner of selectional restrictions rather than selectional preferences (Resnik, 1993). $D^H$ was prepared independently from corpus data. Thus, it contains simple frames having no counterparts in $\tilde{D}$ (and
Figure 2: Frequencies of schemata from $\mathcal{D}^H$ in $\mathcal{D}$

$\text{SEM}K\text{IPI}$), because of sparseness of data. On the other hand, due to data processing errors of $\text{SEM}K\text{IPI}$ (Hajnicz, 2009d; Hajnicz, 2009c), some frames from $\mathcal{D}$ are absent in $\mathcal{D}^H$.

The results were validated w.r.t. a small manually prepared semantic dictionary $\mathcal{D}^H$ composed of all syntactic schemata and corresponding compound semantic frames for 5 verbs: interesować (to interest: 3 schemata), minąć (to pass: 5 schemata), proponować (propose 10 schemata), rozpoząć (to begin: 8 schemata) and widzieć (to see: 13 schemata), which gives total number of 39 schemata. These verbs were selected from the set of 32 ones considered in $\text{SEM}K\text{IPI}$ in a manner maximising their syntactic diversity. The frequency was not a criterion for this choice. However, since the process of aggregation is performed for each syntactic schema separately, their frequency is more important to validate the process.

We should also remember that the task complexity depends on the number of NPs/PrepNPs in the schema. In $\mathcal{D}^H$ there are 12 schemata with 1 NP/PrepNP, 19 schemata with 2 NPs/PrepNPs and 8 schemata with 3 NP/PrepNP. Their frequencies in $\mathcal{D}$ are given in Figure 2. The Figure shows that frequencies of schemata are sufficiently differentiated.

5.2 Validation

There exist three popular clustering validation methods based on co-occurrence of two elements (simple frames) in two partitions of a particular data set. Let

- $b$ be the number of pairs co-occurring in both sets,
- $c$ be the number of pairs co-occurring only in the validated set ($\mathcal{D}$),
- $g$ be the number of pairs co-occurring only in the gold standard ($\mathcal{D}^H$),
- $n$ be the number of pairs co-occurring in neither of sets.

Then Rand statistics ($R$), Jaccard coefficient ($J$) and Folkes and Mallows index ($FM$) are given by the equations (Halkidi et al., 2001):

$$R = \frac{b + n}{b + c + g + n},$$
$$J = \frac{b}{b + c + g},$$
$$FM = \frac{b}{\sqrt{b + c} \sqrt{b + g}}.$$

Rand statistics resemble in a way accuracy measure used in typical lexical acquisition tasks. With such point of view, Jaccard Coefficient and Folkes and Mallows index could be interpret as counterparts of combinations of precision and recall.

In order to apply them to our data (and $\mathcal{D}^H$), we need to bear in mind the specificity of the problem of aggregating semantic dictionary. First, instead of a one large set of data we have plenty of verb syntactic schemata, which frames are aggregated separately. Their validation may be calculated cumulatively or in average. Moreover, there exist some “lonely” frames properly not aggregated with any other frames. In order to take into account such frames (single-element clusters) we consider obvious co-occurrence with itself. Next, the partitioned data sets are different (even though overlapping). Because of that we have counted the above indexes both for all simple frames ($\bigcup$) and for the ones belonging to both dictionaries ($\bigcap$).
The small size of "unusual", and hence they harder to agglomerate. Simple frames belonging only to Dral" ones and hence they are easier to agglomerate. The possible reasons for this could be errors in the similarity measure definition or in the preprocessing, we applied the algorithms to DH distributed back to protodictionary. The results of validation for this case are denoted in Table 1 as hand. The superiority of the agglomerative method is in this case even more apparent.

The fact that the results are better for frames belonging to both dictionaries than for frames belonging to any of them, which is the obvious consequence of the indexes being used: a frame belonging only to one dictionary cannot co-occur with any frame in the second dictionary.

The improvement of Rand statistics calculated cumulatively w.r.t. the one calculated in average indicates the influence of a proportionally large value of n for large schemata. The deterioration of Jaccard coefficient and Folkes and Mallows index calculated cumulatively w.r.t. the one calculated in average indicates the influence of a proportionally large values of c and g. Observe that the larger indexes are the smaller is the difference and the larger is the difference between cumulative and average method of calculating them.

The deterioration of Jaccard coefficient and Folkes and Mallows index calculated cumulatively w.r.t. the one calculated in average indicates the influence of a proportionally large values of c and g. Observe that the larger indexes are the smaller is the difference between cumulative and average method of calculating them.

In order to validate the actual methods without any influence of the corpus preprocessing, we applied the algorithms to DH distributed back to protodictionary. The results of validation for this case are denoted in Table 1 as hand. The superiority of the agglomerative method is in this case even more apparent.

The fact that the results are better for agglomerating D calculated for intersection of dictionary than for redistributed and re-agglomerated DH is a bit surprising. This makes an impression that false simple frames help to agglomerate proper ones. The possible reasons for this could be errors in the similarity measure definition or in the preparation of DH. However, the most probable explanation of this fact is that simple frames belonging to both dictionaries are most "obvious", "natural" ones and hence they are easier to agglomerate. Simple frames belonging only to DH are rare and "unusual", and hence they harder to agglomerate. The small size of DH could influence the results as well.

---

### Table 1: Validation of aggregation of frames

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>J</th>
<th>FM</th>
<th>R</th>
<th>J</th>
<th>FM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggl</td>
<td>77.6</td>
<td>26.7</td>
<td>40.1</td>
<td>83.6</td>
<td>9.7</td>
<td>17.8</td>
</tr>
<tr>
<td>MST-80</td>
<td>73.4</td>
<td>22.3</td>
<td>35.2</td>
<td>79.3</td>
<td>5.6</td>
<td>10.8</td>
</tr>
<tr>
<td>MST-90</td>
<td>63.6</td>
<td>19.5</td>
<td>30.1</td>
<td>67.7</td>
<td>2.8</td>
<td>8.1</td>
</tr>
<tr>
<td>Aggl</td>
<td>91.3</td>
<td>86.4</td>
<td>91.8</td>
<td>82.8</td>
<td>69.9</td>
<td>82.3</td>
</tr>
<tr>
<td>MST-80</td>
<td>87.9</td>
<td>82.6</td>
<td>89.1</td>
<td>77.7</td>
<td>59.7</td>
<td>75.0</td>
</tr>
<tr>
<td>MST-90</td>
<td>83.3</td>
<td>78.0</td>
<td>86.5</td>
<td>66.5</td>
<td>55.0</td>
<td>73.9</td>
</tr>
<tr>
<td>Aggl</td>
<td>87.5</td>
<td>73.3</td>
<td>82.9</td>
<td>92.6</td>
<td>68.8</td>
<td>81.5</td>
</tr>
<tr>
<td>MST-80</td>
<td>75.2</td>
<td>51.4</td>
<td>66.8</td>
<td>82.9</td>
<td>17.2</td>
<td>39.4</td>
</tr>
<tr>
<td>MST-90</td>
<td>77.8</td>
<td>57.2</td>
<td>71.5</td>
<td>83.3</td>
<td>20.9</td>
<td>42.7</td>
</tr>
</tbody>
</table>

The results of the validation are presented in Table 1. They show that the best results are obtained for the agglomerative method. The results are mostly better for frames belonging to both dictionaries than for frames belonging to any of them, which is the obvious consequence of the indexes being used: a frame belonging only to one dictionary cannot co-occur with any frame in the second dictionary.

6 Conclusions

In this paper two methods of aggregating simple semantic frames into semantically coherent compound ones were discussed and compared. The fact that a simple agglomerative method was better than MST is indication to apply more sophisticated agglomerative methods.

We also plan to extend DH, which will enable us to perform the more reliable validation. In particular, the validation w.r.t. the number of NPs/PrepNPs in a schema and/or the number of simple frames in it will be possible, which is disabled by the present small size of DH.

---

**References**


