Similarity measure between frames for Polish semantic valence dictionary

Elżbieta Hajnicz

Institute of Computer Science, Polish Academy of Sciences
ul. Ordona 21, 01-237 Warsaw, Poland
Elzbieta.Hajnicz@ipipan.waw.pl

Abstract

In this paper we define a similarity measure between semantic valence frames. It is based on a similarity measure between senses of verb arguments. Senses are represented by general semantic categories or wordnet synset. The measure is supposed to be used for aggregation of semantic valence frames that differ in NP/PP slot senses but represent the same meaning of the verb.

Keywords: semantic similarity, wordnet, semantic valence dictionary, Polish

1. Introduction

The main goal of our work is to enrich a valence dictionary of Polish verbs by adding semantic information. This information may be represented by means of general wordnet semantic categories of nouns (cf. Table 1) or by synsets from the entire net. The plain syntactic valence dictionary is a collection of predicates (here: verbs) provided with a set of verb frames. Verb frames consist of syntactic slots that represent phrases occurring in the corresponding position in a sentence. Thus, our goal is to provide syntactic slots (here: NPs/PPs) with a list of appropriate semantic categories for the corresponding nouns.

For Polish, any hand-crafted semantic valence dictionary, such as VerbaLex (Hlaváčková and Horák, 2006) for Czech, does not exist. Our work consists in creating such a resource on the basis of a semantically annotated corpus of texts. In (Hajnicz, 2009c) the process of semantic annotation of nouns which are semantic heads of NPs/PPs by means of PlWN semantic categories (see section 2.) is discussed. In (Hajnicz, 2009b) its adaptation to the entire PlWN hierarchy of hypernymy is presented. In (Hajnicz, 2009a) the process of gathering information from the annotated corpus into a dictionary is shown.

However, a dictionary obtained in such a way has a plenty of entries, with a single sense (a semantic category or a synset) assigned to each syntactic slot. This does not reflect the actual semantics of a verb, since different senses of arguments do not entail different meaning of a verb. In other words, such categorisation is too fine-grained. For instance in sentences,

Piotr$_{\mathsf{person}}$ *przejechał park*$_{\mathsf{location}}$ *brata samochodem*$_{\mathsf{artifact}}$.
(*Piotr crossed his brother's park in a car.*)

Piotr$_{\mathsf{person}}$ *przejechał psa*$_{\mathsf{animal}}$ *brata samochodem*$_{\mathsf{artifact}}$.
(*Piotr run over his brother's dog by a car.*)

We have different meanings of the verb *przejechać*. These differences appears 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 valence dictionary, with location and animal on the object position, respectively. On the other hand, in the sentences:

Piotr$_{\mathsf{person}}$ *kupił bratu*$_{\mathsf{person}}$ *park*$_{\mathsf{location}}$.
(*Piotr bought his brother a park.*)

Piotr$_{\mathsf{person}}$ *kupił bratu*$_{\mathsf{person}}$ *psa*$_{\mathsf{animal}}$.
(*Piotr bought his brother a dog.*)

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 need a similarity measure between senses. It is based on the assumption that two senses are put together only if all senses positioned in between by means of a particular similarity measure are senses of a considered slot. Observe that one can buy almost everything, in particular things having semantic categories positioned in between animal and location (cf. Figure 1). Contrary, objects of *crossing* and *running over* are separated.

2. Data resources

Our main resource for defining a similarity measure is a wordnet. Since we work on Polish data, we use the Polish WordNet (Derwojedowa et al., 2007, 2008a,b), called *Słowosieć* (English acronym PlWN). PlWN is a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet (Fellbaum, 1998) and those constructed in the EuroWordNet project (Vossen, 1998). Polish WordNet describes the meaning of a lexical unit (LU) of one or more words by placing this unit in a network which represents such relations as synonymy, hypernymy, meronymy, etc. For the construction of a semantic valence dictionary we need only the part of

category	top synsets	category	top synsets	category	top synsets
act	action	communication	code	phenomenon	nature
	activity		communication		phenomenon
animal	animal		etnolect	plant	plant
artifact	concept		information		plant part
	medium		sound	possession	business
	net	event	event		economics
	thing	feeling	feeling		possession
attribute	detail	food	drink	process	process
	feature		food	quantity	degree
body	body	group	fauna		quantity
	corpse		flora		value
	gene		group	relation	connection
	meat		set		domain
	organ	location	container		institution
	tissue		line		relation
cognition	cognition		part		team
	destiny		place	shape	shape
	exercise		space	state	state
	idea	motive	duty	substance	substance
	intelligence		goal		sbst state
	psyche		position		sbst type
			solution	time	course
		object	object		period
		person	person		phase

Table 1: The set of tops in the hypernymy hierarchy correlated with the predefined set of semantic categories in Polish WordNet

PlWN describing relations between nouns. We have focused on the synonymy represented by synset classes and on hypernymy represented by a directed acyclic graph. Synsets having no hypernyms (positioned on the top of the hypernymy hierarchy) are called *top synsets*.

Before the entire PlWN net was constructed, the set of LUs was divided according to the predefined set of 25 general *semantic categories*. The process of the manual creation of the net was performed to the large extent w.r.t. this division. As a result, each synset contains LUs of the same category. To some extent this concerns hypernymy as well. In table 1 the list of LUs representing top synsets are presented together with the corresponding semantic categories.

Observe that some of the semantic categories are extremely heterogeneous. The majority of nouns which are categorised as group are groups of people, but similar categorisation have groups of animals, plants and even sets of things. Plurality is of course an important feature of beings, but from the semantic valence dictionary creation point of view more important is whether we deal with people, animals etc.

A syntactic dictionary is a list of entries representing sets of syntactic schemata appropriate for a particular verb. Every schema is composed of the lemma of a verb and a list of its arguments. The list of arguments can include: adjective phrases AdjP, adverb phrases AdvP, infinitive phrases InfP, noun phrases NP, prepositional-adjective phrases PAP, prepositional-nominative phrases PP, clauses SentP and reflexive marker *się*.

Arguments can be parametrised. AdvP has no parameters, the only parameter of AdjP and NP is their case, the only parameter of InfP is its aspect. PAP and PP have two parameters: a preposition and the case of its AdjP or NP complement, respectively. SentP has one parameter showing type of a clause.

Below we list syntactic dictionary entries for the verb *proponować* (*to propose*).

(1) proponować infp np:dat np:nom
proponować infp np:nom
proponować np:acc np:dat np:nom
proponować np:acc np:nom prepnp:dla:gen
proponować np:acc np:nom prepnp:na:acc
proponować np:acc np:nom żeby
proponować np:dat np:nom żeby
proponować np:nom pz
proponować np:nom że
proponować np:nom żeby

Formally, a syntactic valence dictionary $\mathcal{D}$ is a set of pairs $\langle v, g\rangle$, where $v \in V$ is a verb and $g \in G$ is a syntactic frame. A semantic dictionary $\mathfrak{D}$ is a set of tuples $\langle\langle v, g, f\rangle, n, m\rangle$, where $\langle v, g\rangle \in \mathcal{D}$ is a schema of a verb, $f \in F_g$ is one of its frames, n is the frequency of $\langle v, g\rangle$ and m is the frequency of $\langle v, g, f\rangle$. Furthermore, $g = \langle s_1, \ldots, s_n\rangle$, where $s_i \in S$ are syntactic slots, and $f = \langle\langle s_1, c_1\rangle, \ldots, \langle s_n, c_n\rangle\rangle$, where $c \in C$ is a sense.

A semantic valence dictionary of 32 verbs was collected both for semantic categories (Hajnicz, 2009a) and synsets, basing on the corresponding semantic annotation of nouns (Hajnicz, 2009c,b). In (2) we present a part of an exemplary entry for the schema `np:acc np:dat np:nom` of the verb *proponować* (*to propose*) based on semantic categories' annotation.

(2) proponować

np:acc np:dat np:nom	573
acc: act; dat: act; nom: group	8.63
acc: act; dat: act; nom: person	31.29
acc: act; dat: feeling; nom: group	4.63
acc: act; dat: group; nom: event	1.00
acc: act; dat: group; nom: group	16.63
acc: act; dat: group; nom: person	50.35
acc: act; dat: locat.; nom: person	9.29
acc: act; dat: person; nom: artif.;	5.21
acc: act; dat: person; nom: group	22.63
acc: act; dat: person; nom: person	51.36
acc: act; dat: person; nom: quant.	0.83
acc: act; dat: quant.; nom: group	5.63
acc: attrib.; dat: person; nom: group	0.50
acc: attrib.; dat: person; nom: person	2.08
acc: feeling; dat: person; nom: person	1.58
acc: time; dat: group; nom: person	2.44
acc: time; dat: person; nom: person	3.45

It is obvious that a person or a group (`np:nom`) can propose almost anything (`np:acc`) to a person or a group (`np:dat`) and that corresponding source frames should be aggregated into one resulting semantic frame. Institutions are often categorised as relations, which justifies their occurrence on the nominative and dative positions. Categories gathered as their accusative complements are act, artifact, attribute,[1] cognition, communication, event,feeling, location, person, possession, relation, state and time.

Other frames will be probably aggregated separately. Some of them result from the less obvious use of a verb. For instance, *gazety* (*newspapers*), categorised as artifact, or *reklamy* (*adverts*) can propose something to their readers. Other frames are effects of errors appearing both in syntactic and semantic selection process. Some of them will be deleted after pruning.

3. Similarity measure

Any algorithm for frames aggregation[2] needs a similarity measure to estimate the similarity between frames. Most popular statistical clustering algorithms of this type are k-Means (MacQueen, 1967) and MST (Gower and Ross, 1969).

The main idea is that the similarity between two frames is a function of the similarity between senses for particular slots.

3.1. Similarity between senses

There exists a number of similarity measures between word meanings based on wordnet hypernymy hierarchy (Budanitsky and Hirst, 2006). However, they concern synsets located in the similar area of the hierarchy and do not concern tops of the hierarchy. Contrary, we are interested only in similarity between tops. An important assumption made for a dictionary creation is that if a synset is not a top one, then the frame should not be aggregated for the slot it appears in.

Another important assumption is that we deal with one measure for all verbs and all syntactic slots senses are assigned to.

In (Hajnicz and Wiech, 2008) a method of determining similarity between semantic categories based on triples ⟨verb, slot, category⟩ extracted from an annotated corpus is presented. Unfortunately, the resultant similarity measure is based on a linear order. However, we have a deep impression that the ordering of senses is not linear.

Finally, we decided to determine this measure manually. The similarity measure between semantic categories is presented in Figure 1, whereas the similarity measure between PLWN top synsets is presented in Figure 2. Such a measure will be denoted as $d\colon C \times C \longrightarrow \mathbb{R}^+$. Senses are represented as nodes of a graph and distances between them are represented as labels of arcs linking them. For all pairs of non-adjacent senses in a graph, the distance between them is calculated as the length of the shortest path between them. Please note that the graphical composition of pictures 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.

The similarity measure between semantic categories was drown up before the entire PLWN structure was established on the basis of categorisation of most frequent nouns and intuition. The heterogeneity of some semantic categories caused some simplifications in it. Observe that top synsets having the same category assigned are often separated by the corresponding similarity measure of Figure 2. This is another consequence of heterogeneity of a category (e.g., group, relation, possession).

The main idea behind the measures construction was to keep together material beings, immaterial beings and time-dependent entities. This idea was accomplished to a great extent. Some decisions concerning distances between top synsets are better understandable when one knows their hyponyms.

[1] E.g., noun *swoboda* used as an object of proposition is categorised as attribute.

[2] We do not use the popular term *clustering* used for similar tasks, since the result of aggregation is a single dictionary entry, not a collection or a class.

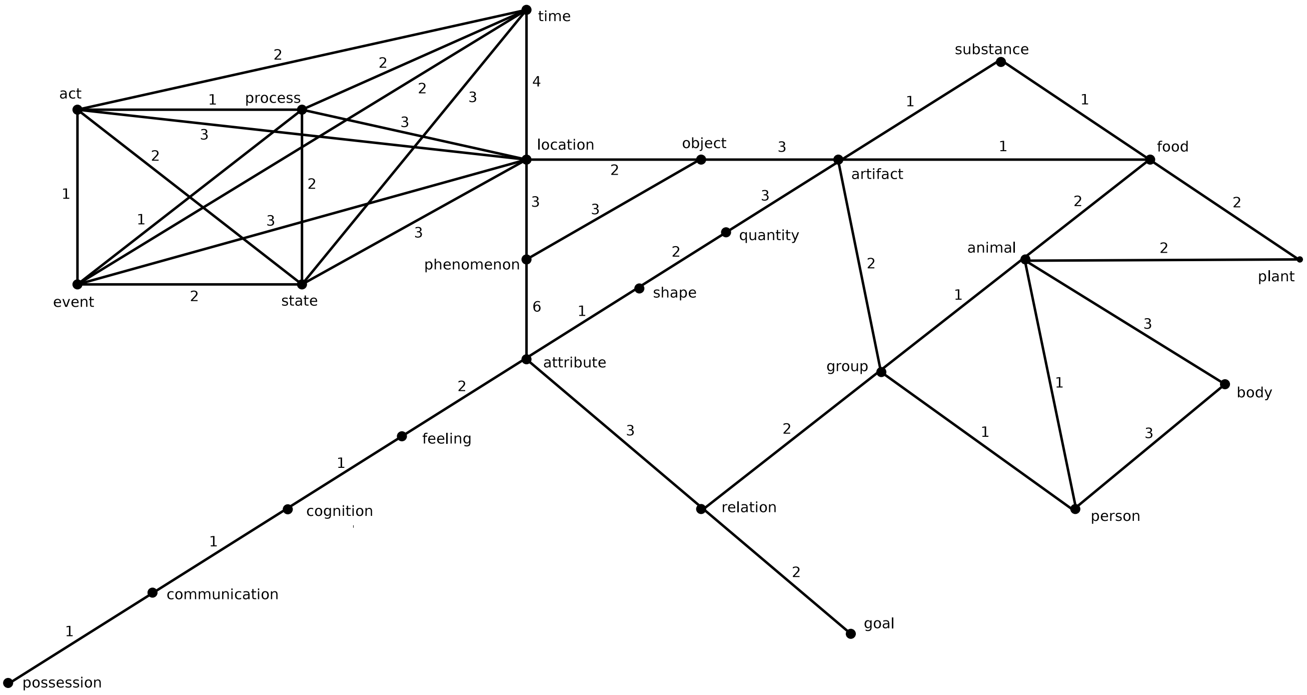

Figure 1: Similarity measure between semantic categories

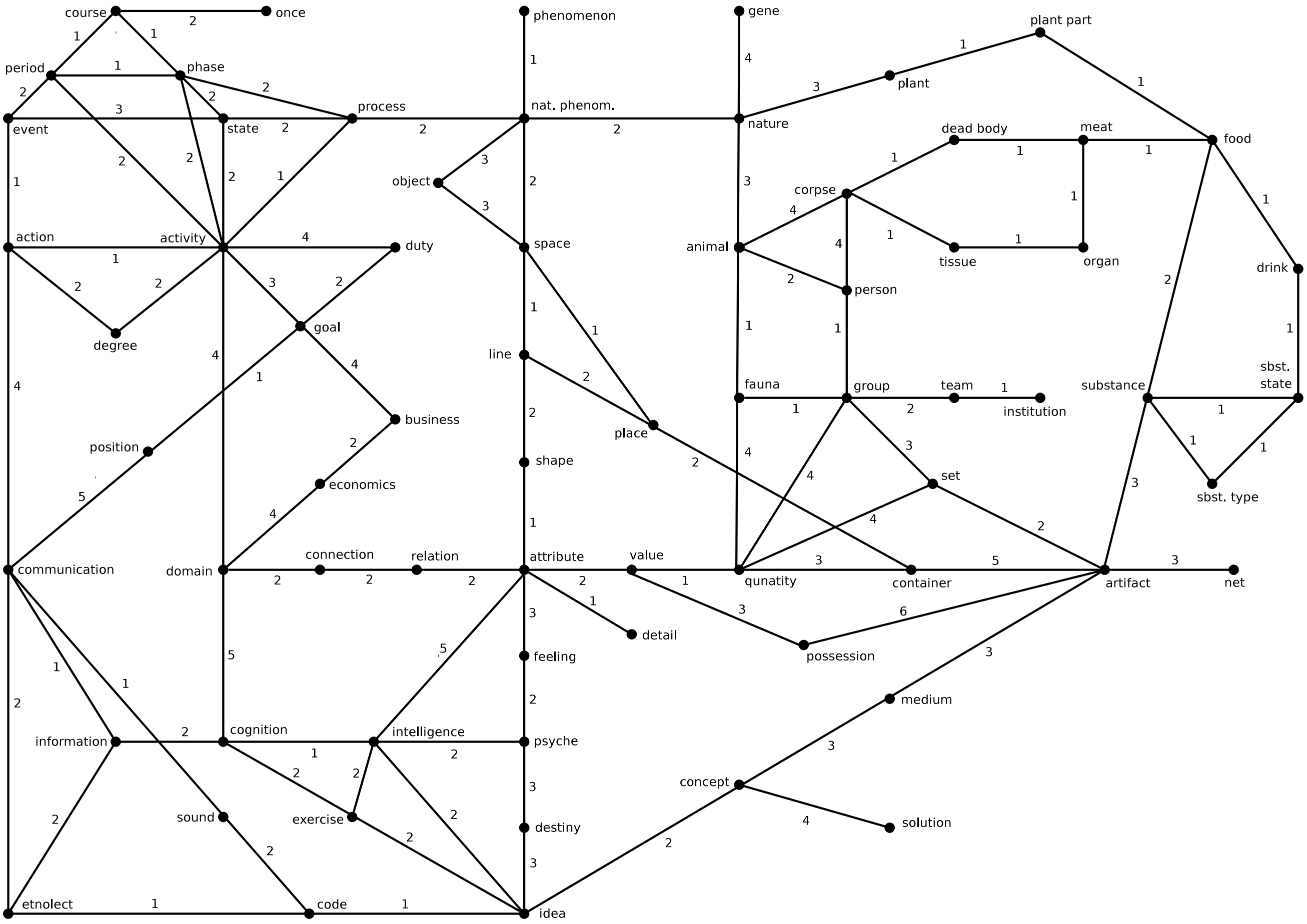

Figure 2: Similarity measure between top synsets

3.2. Similarity between frames

A similarity measure between frames D is defined on the basis of the similarity measure between senses d. This definition is general and does not depend on a particular choice of d. Frames to compare represent the same syntactic schema, and in particular they have the same number of arguments n. The idea is to join similarity measures d_i counted for each slot i, hence we deal with a set of measures D^n.

Definition 1. Let $f^1 = \langle\langle s_1, c_1^1\rangle, \ldots, \langle s_n, c_n^1\rangle\rangle$ and $f^2 = \langle\langle s_1, c_1^2\rangle, \ldots, \langle s_n, c_n^2\rangle\rangle$ be frames of the same syntactic schema $g == \langle s_1, \ldots, s_n\rangle$ (i.e., $\langle\langle v, g, f^1\rangle, n, m^1\rangle, \langle\langle v, g, f^2\rangle, n, m^2\rangle \in \mathfrak{D}$). We defined $D^n \colon C^n \times C^n \longrightarrow \mathbb{R}^+$ in the following way:

$$D^n(f^1, f^2) = \sqrt{\frac{\left(d(c_1^1, c_1^2)\right)^2 + \cdots + \left(d(c_n^1, c_n^2)\right)^2}{\varphi(n)}}.$$

The numerator of the fraction is a standard Euclidean measure. Its denominator is a constant weighting the measure. Standard value of $\varphi(n)$ is 1. However, also $\varphi(n) = \sqrt(n)$ or even $\varphi(n) = n$ is allowed in order to make D^n independent from the length of a schema n. The Euclidean character of D^n entails the following proposition:

Proposition 1. If d is a metrix, then D^n is a metrix for each n.

4. Conclusions

In this paper a method of defining a similarity measure between semantic valence frames was proposed. It is based on a similarity measure between senses of nouns being heads of verb arguments. The definition is Euclidean and the choice of a measure between senses is free. We assumed that we have one measure for all slots, but the definition could be easily generalised to the case when there exist different similarity measures for various slots.

The measures enable us to perform an aggregation of semantic valence dictionary. Evaluation of this process would serve for evaluation of the measures themselves. This is a task for further work.

Acknowledgements

This paper is a scientific work supported within the Ministry of Science and Education project No N N516 0165 33

References

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):1–27.

Derwojedowa, M., Piasecki, M., Szpakowicz, St. and Zawisławska, M. (2007). Polish WordNet on a shoestring. In: *Data Structures for Linguistic Resources and Applications: Proceedings of the GLDV 2007 Biannual Conference of the Society for Computational Linguistics and Language Technology*. Universität Tübingen, Tübingen, Germany.

Derwojedowa, M., Piasecki, M., Szpakowicz, St., Zawisławska, M. and Broda, B. (2008a). Words, concepts and relations in the construction of Polish WordNet. In: Tanacs, A., Csendes, D., Vincze, V., Fellbaum, Ch., Vossen, P. (Eds.) *Proceedings of the Global WordNet Conference*. Seged, Hungary.

Derwojedowa, M., Szpakowicz, St., Zawisławska, M. and Piasecki, M. (2008b). Lexical units as the centrepiece of a wordnet. In: Kłopotek et al. (2008).

Fellbaum, Ch. (Ed.) (1998). *WordNet — An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gower, J.C. and Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18(1):54–64.

Hajnicz, E. (2009a). Generalizing the em-based semantic category annotation of NP/PP heads to wordnet synsets. in this volume.

Hajnicz, E. (2009b). Problems with pruning in automatic creation of semantic valence dictionary for Polish. In: Matoušek V., Mautner, P. (Eds.) *Proceedings of the International Conference on Text, Speech and Dialogue TSD 2009*, volume 5729 of *LNAI*. Berlin, Heidelberg: Springer-Verlag.

Hajnicz, E. (2009c). Semantic annotation of verb arguments in shallow parsed Polish sentences by means of EM selection algorithm. In: Marciniak M., Mykowiecka, A. (Eds.) *Aspects of Natural Language Processing*, volume 5070 of *LNCS*. Springer-Verlag, pp. 211–240.

Hajnicz, E. and Wiech, M. (2008). Applying grade methods to detect similarity of semantic categories of nouns for semantic valence dictionary creation. In: Kłopotek et al. (2008), pp. 259–268.

Hlaváčková, D. and Horák, A. (2006). Verbalex — new comprehensive lexicon of verb valences for Czech. In: *Proceedings of the Third International Seminar on Computer Treatment of Slavic and East European Languages*. Bratislava, Slovakia.

Kłopotek, M.A., Przepiórkowski, A. and Wierzchoń, S.T. (Eds.) (2008). *Proceedings of the Intelligent Information Systems XVI (IIS'08)*, Challenging Problems in Science: Computer Science. Zakopane, Poland: Academic Publishing House Exit.

MacQueen, J.B., (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5-th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press.

Vossen, P. (Ed.) (1998). *EuroWordNet: a multilingual database with lexical semantic network*. Dordrecht, Holland: Kluwer Academic Publishers.