Optimizing semantic granularity for NLP - report on a lexicographic experiment[1]

Silvie Cinková, Martin Holub & Vincent Kríž

Keywords: *corpus pattern analysis, semantic tagging, semantic granularity, English, verbs.*

Abstract

Experiments with semantic annotation based on the Corpus pattern Analysis and the lexical resource PDEV (Hanks and Pustejovsky, 2005), revealed a need of an evaluation measure that would identify the optimum relation between the semantic granularity of the semantic categories in the description of a verb and the reliability of the annotation expressed by the interannotator agreement (IAA). We have introduced the Reliable Information Gain (RG), which computes this relation for each tag selected by the annotators and relates it to the entry as a whole, suggesting merges of unreliable tags whenever it would increase the information gain of the entire tagset (the number of semantic categories in an entry). The merges suggested in our 19-verb sample correspond with common sense. One of the possible applications of this measure is quality management of the entries in a lexical resource.

1. Introduction

The "semantic granularity" of word senses (readings, categories, lexical units) in the entries of a lexical resource is of great importance for any automatic lexical semantic analysis. The most traditional discipline in this field is word sense disambiguation (WSD). Although the two tacit assumptions under which WSD has been pursued, namely that 1) various uses of polysemous words can be sorted into discrete senses and 2) the senses reflect a mental reflection of the given word shared by the entire language community, appear to be falsified by the generally low interannotator agreement (IAA) in WSD tasks, the existing major lexical resources (Fellbaum, 1998, Ruppenhofer et al., 2010, Palmer et al., 2005 and Weischedel et al., 2011) are still maintained and their annotation schemes are adopted for creating new manually annotated data. More to say, they are not only used in WSD and semantic labeling, but also in more recent research directions that in their turn do not rely on an inventory of discrete word senses any more, e.g. in distributional semantics (Erk, 2010) and textual entailment recognition (Zanzotto et al., 2009 and Aharon et al., 2010).

The synsets in WordNet have been reflected as too granular and thus impossible to disambiguate in confrontation with actual language data, which results in poor IAA and low reliability of the manually annotated data for machine-learning tasks. PropBank and OntoNotes seek to increase the reliability of the annotation by increasing IAA. IAA around 90% is reached in both cases by making the senses more coarse-grained. The data are therefore more reliable, but the entries deliver less information.

To the best of our knowledge, there has been no measure in practical use that would relate granularity, reliability of the annotation (derived from IAA) and the resulting information gain. We have formulated such a measure – **"**RG" (reliable information gain**,** Cinková et al. to appear1**)** - and we have implemented it in an algorithm that analyzes a lexicon entry considering the IAA from three annotators measured on a 50-sentence random corpus sample. The algorithm computes the information gain for each sense and suggests possible optimizations (merging).

While normally the usefulness of a lexical resource can hardly be assessed before it is used in an application (which in turn hardly happens before the resource is finished), using this algorithm helps managing the optimum level of semantic granularity for each finished entry.

2. Interannotator agreement experiments on PDEV and VPS-30-En

2.1. *Motivation*

The idea of RG has crystallized as a supplementary result of a two-year cooperation with Patrick Hanks at the Institute of Formal and Applied Linguistics, Charles University in Prague. We were exploring the Pattern Dictionary of English Verbs (Hanks and Pustejovsky, 2005) from the point of its usability in NLP tasks, since the patterns of PDEV appear intuitively very appealing for automatic clustering of semantically similar verb uses by means of statistical machine learning. As a first step, we considered it necessary to find out whether the clustering provided by P. Hanks can be replicated by other people, assuming they are familiar with the clustering principle.

2.2. *PDEV vs. other lexical resources*

The principle behind PDEV is different from WordNet and OntoNotes word senses. The Theory of Norms and Exploitations and the method of Corpus Pattern Analysis, on which PDEV draws (Hanks, forthcoming), are an example of the modern, corpus-based lexicology that has boomed since the 1990s (Sinclair, 1991, Fillmore and Atkins, 1994, Kilgarriff, 1997 and many more) and has had a great impact on the practical lexicography. There has been a general consensus that dictionary definitions need to be supported by corpus examples. The lexical description in modern English monolingual dictionaries (Sinclair et al., 1987, Rundell et al., 2007) explicitly emphasizes contextual clues, such as typical collocates and the syntactic surroundings of the given lexical item, rather than relying on very detailed definitions. In other words, the sense definitions are obtained as syntactico-semantic abstractions of manually clustered corpus concordances: in classical dictionaries as well as in lexical resources for NLP.

Being one of the leading researchers in this direction, Hanks has been performing a radical revision of the lexical description both in theory (Hanks, forthcoming) and in practice (Hanks and Pustejovsky, 2005), creating the Pattern Dictionary of English Verbs (PDEV), publicly available at http://nlp.fi.muni.cz/projekty/cpa/). Instead of an inventory of senses, the verb is supposed to manifest itself in different corpus concordances, which activate different aspects of its “meaning potential”. The verb entries consist of “categories”. Each category consists of a “pattern” of regular use and an “implicature”. The pattern has the form of a proposition, and so does the implicature, which can be regarded as a paraphrase of the pattern. The categories define, roughly speaking, normal uses of a verb that have a common syntactic, lexical and morphological features. They denote a similar event in which similar participants (e.g. humans, artifacts, institutions, and other events) are involved.

During the entry compilation process, several hundred random BNC concordances (The British National Corpus, 2007) are manually clustered according to the syntactic, lexical and morphological similarity, as well as according to the semantic similarity of the implicatures. The patterns constitute “prototypes” (Hanks, forthcoming). Concordances that match these prototypes well are called “norms”. Concordances that match with a reservation are called “exploitations”. The corpus annotation of PDEV indicates the norm-exploitation status for each concordance.

The main strength we see in PDEV is the tight link between each category and the concordances it is based on. Rumshisky et al. (2009) conducted an interesting experiment, which suggested that better IAA in lexical semantic analysis could be reached if the classical WSD task was redesigned: her annotators, who did not even need to be linguists, reached a pretty good agreement on to what extent a verb is used in the same sense in two different concordances. The experimental data was taken from PDEV. The resulting clusters were somewhat coarser-grained than the original PDEV categories, since the annotators had not been instructed to consider the syntactic features of the concordances, such as the argument structure. Nevertheless, they displayed a significant overlap with the categories. This raised the hope that linguists could reach a reasonable IAA even on the original PDEV categories.

2.3. *The annotation experiments*

The original PDEV had never been tested on IAA. Each entry had been based on concordances annotated solely by the author of that particular entry. The annotation instructions had been transmitted only orally. The data had been evolving along with the method, which implied that the concordance sample was not always completely revised after a revision of the entry and that the clustering criteria were different for different entries, for instance: how much surrounding context and common knowledge should be taken into account and which is the stronger criterion – the lexical, morphological and syntactic similarity of the pattern or the semantic similarity of the implicature.

We ran several annotation rounds in different setups. In 2009, we appointed several annotators. All were linguists and one of them was P. Hanks himself. We picked 20 finished entries with annotated concordances, presented them for the annotators and asked them to sort another set of 50 new random concordances for each verb in the same way. The verbs were *abstain, accept, address, admit, alter, announce, argue, call, claim, engage, explain, fire, lead, need, plan, rush, say, spoil, tell* and *visit*. The interannotator agreement (IAA, measured by Fleiss’ kappa (Artstein and Poesio, 2008)) was shifting for each verb. For instance, eleven verbs reached over 0.6, which is generally considered a fair agreement. Three verbs reached over 0.7, but other three not even 0.3. The average was 0.596. The manual analysis of interannotator disagreements resulted in a preliminary annotation guide (Cinková and Hanks, 2010), which preserved a snapshot of the CPA theory and also contained some tentative suggestions for the next annotation experiments, such as an elaboration on the exploitations. We divided the explotations into four types: figurative use, unexpected lexical population of an argument, syntactic deviation and coercion.

Later on, it became evident that we would need an experimental sandbox on our own not to destroy the original PDEV. In late 2010, we parted from PDEV and started building **VPS-30-En** (Cinková et al., to appear2)**.**

2.4. *VPS-30-En*

VPS-30-En (Verb Pattern Sample of 30 English verbs) is a collection of 30 revised PDEV verbs in which the adjustments of the entries and the original concordance samples were driven by IAA findings, with the optimization of RG in mind. It contains the verbs *access, ally, arrive, breathe, claim, cool, crush, cry, deny, enlarge, enlist, forge, furnish, hail, halt, part, plough, plug, pour, say, smash, smell, steer, submit, swell, tell, throw, trouble, wake* and *yield*.

The collection is not a competitor of PDEV, but a deliberately small sample revised and cleaned up as a gold-standard data set for statistical pattern recognition. VPS-30-En can be browsed and downloaded at http://ufal.mff.cuni.cz/spr and soon also in the LINDAT-CLARIN repository.

We revised the 30 verbs and ran the annotations in several rounds according to the new guidelines and with different annotators. We have also slightly altered the annotation scheme, revised some entries and updated the reference samples (usually 250 concordances per verb). The annotators were given the entries as well as the reference sample annotated by the lexicographer (S. Cinková) and a test sample of 50 random concordances for annotation. We measured IAA and analyzed the interannotator disagreement manually. When the IAA was low and the type of disagreement indicated a problem in the entry, the entry was revised again. Then the lexicographer revised the original reference sample along with the first 50-concordance sample. The annotators got back the revised entry, the newly revised reference sample and an entirely new 50-concordance annotation batch. The final multiple 50-concordance sample was subject to "adjudication": the lexicographer compared the three annotations and eliminated evident misjudgments. The adjudication protocol has been kept for further experiments. In the end, we got for each verb an entry along with 300+ manually annotated concordances (single values), out of which 50 are manually annotated and adjudicated concordances (multiple values cleared of evident misjudgments).

The analysis of the annotator disagreements made it clear for us that each verb has its individual maximum IAA, at least with respect to the Fleiss' kappa. Fleiss' kappa, unlike, for example, the sheer percentage count, considers the "perplexity" of each verb. A verb's perplexity is the higher, the more tags (i.e. pattern numbers) are used and the more evenly they are distributed. The perplexity drops when the assigned pattern numbers are few and when one or a few tags dominate. Fleiss' kappa drops more significantly with each interannotator disagreement in low-perplexity verbs than in high-perplexity verbs. A distinct case of a low-perplexity verb is the verb *halt* in Fig. 1, which has only three categories and one is dominant. The 4 annotators agreed to approx. 80%, but the Fleiss' kappa did not even reach 0.6, whereas a highly perplex verb as *part* reached a 0.7 kappa with the same agreement percentage.

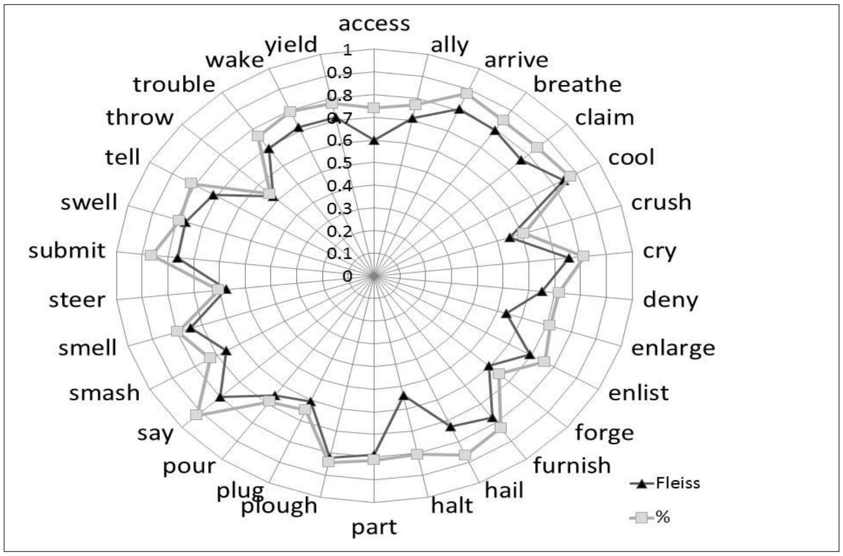

Figure 1. The relation between IAA measured by percentage and by Fleiss' kappa.

3. Information Gain Optimization

3.1. *IAA evaluation*

It is illusory to require a 100% agreement among four annotators in a semantic task on a 50-concordance sample. The average number of categories in an entry is 13.6, but potentially it is five times higher, since the annotators are expected to observe four types of exploitations in addition and we consider even disagreements within the same category number, when the annotators differ in the exploitation markup. When exploitations are considered, the average Fleiss' kappa in VPS-30-En is 0.7. When we neglect them, it is 0.791. What do these numbers tell us about the agreement? How do we learn for each individual entry-sample pair at which point the annotation is reliable enough and the semantic information is granular enough to be interesting for text analysis? As an answer to these questions, we have introduced RG, a measure that optimizes the relation between the informative power of a tag and its reliability with respect to how probably annotators would agree to assign this particular tag.

3.2. *Computing the Reliable Information Gain (RG)*

Confusion matrices for each annotator pair (Fig. 2) are produced for the analysis of each annotated entry. One annotator is represented by the lines and the other by the columns. The numbers inside the matrix indicate how many times a combination of tags from these two annotators occurred. For example, A1 and A2 agreed in 29 concordances on the tag 1. In three concordances A1 assigned tag 5, whereas A2 assigned 4. These matrices constitute the input of the RG optimization algorithm for the computation of "confusion probability matrices" (CPM, Fig. 3**Error! Reference source not found.**): each row in the CPM corresponds to one particular tag and shows the probabilities of selecting possible tags by different annotators **on the assumption that** one of the annotators has already chosen the given tag. For example, the first row indicates that **if** one annotator uses the tag 1, there is an 89.5% chance that another annotator will agree, an 8.4% chance that another annotator will choose the tag 1.a, and a 2.1% chance that another annotator will choose the tag 2.

A_1 vs. A_2						A_1 vs. A_3						A_2 vs. A_3					
	1	1.a	2	4	5		1	1.a	2	4	5		1	1.a	2	4	5
1	29	1	1	0	0	1	29	2	0	0	0	1	27	2	0	0	0
1.a	0	1	0	0	0	1.a	1	0	0	0	0	1.a	2	0	1	0	0
2	0	1	11	0	0	2	0	0	12	0	0	2	1	0	11	0	0
4	0	0	0	2	0	4	0	0	0	1	1	4	0	0	0	1	4
5	0	0	0	3	1	5	0	0	0	0	4	5	0	0	0	0	1

Figure 2. Confusion matrices for the paired annotators A1, A2 and A3.

This enables us to assess the “usefulness” of each tag (and hence the usefulness of each pattern). The usefulness drops when the tag is unreliable. The usefulness of each tag contributes to the “average reliable gain” (ARG), which is computed for the entire tagset (all patterns). ARG would reach its maximum if there was a 100% IAA. Since in practice the IAA we get is lower, ARG should be maximized by merging some of the tags with poor RG. When there is a lot of confusion among annotators, our algorithm maximizes ARG and suggests the best set of merges to produce an optimized tagset, so that the degree of confusion among tags is reduced and the reliable information is preserved at the same time.

	1	1.a	2	4	5
1	0.895	0.084	0.021	0.000	0.000
1.a	0.727	0.091	0.182	0.000	0.000
2	0.053	0.053	0.895	0.000	0.000
4	0.000	0.000	0.000	0.333	0.667
5	0.000	0.000	0.000	0.571	0.429

Figure 3. A confusion probability matrix.

We have investigated 19 verbs: *access, wake, enlarge, forge, enlist, crush, hail, part, pour, smash, smell, deny, yield, ally, cry, halt, plough, submit* and *arrive*. In all verbs, the algorithm first suggested merging norm tags with exploitation tags of the same pattern (e.g. 5, 5.a, 5.s, 5.f), which is perfectly in accordance with the common sense. It also often suggested merging various types of exploitations with u (“unclassifiable”) or x (“not a verb”). Three phenomena in the annotated concordances have triggered the more substantial merging suggestions: participle uses of verbs, coercion and meaning modulation in verb arguments. The participle form, especially in attributive position or when combined with *be*, obscures the actual number of event participants, since in many verbs the event can be interpreted as passive, pseudo-passive or genuine adjective at the same time. To reduce unnecessary annotator confusion, we have introduced special patterns for participial forms, when the 250-concordance reference sample suggested that they are frequently used. We do not regard it as a serious disagreement when one annotator selects the underspecified participial pattern whereas the other dares to interpret the event in more detail and picks the transitive pattern (selecting intransitive patterns is forbidden for syntactic reasons). Cf.:

(1) […] inmates in prisons , <enlisted> men in basic training camps […]

Enlist in this case fits both the transitive pattern (someone enlists someone else to provide him with a service) and the participial pattern "enlisted = registered for a service". The intransitive pattern "someone enlists (in the army etc.)" is out of question, since we regard the pseudo-passive verb form syntactically as an adjective (and an adjective would get "x").

This helps eliminating a more substantial disagreement between two very different patterns (transitive vs. intransitive). Suggestions of merges between a regular pattern and a participial pattern are therefore systematically rejected as unnecessary. The merging suggestions become, however, interesting, when an argument position of a verb is typically occupied by nouns that enable meaning modulations or coercion. For instance, *support* that *arrives* can be both interpreted as an event that happens and as material entities that are delivered (each belonging to a different pattern), or one can *submit* himself to *the will of others,* where we either emphasize *will* as an (unpleasant) eventuality/rule or the people who execute their will (submit oneself to a person/institution), which also have each their respective patterns. When modulations/coercions occur systematically, the merge is worth consideration.

4. Discussion

A manual analysis of the suggested merges showed that the suggested merges were all in accordance with the common sense. Besides, the algorithm suggested very few merges regarded as substantial. Given that semantic annotation is meant to teach computers to mimic human judgment, we consider it a good preliminary result that the algorithm behaves in a way that makes sense to humans, and, in addition, it verifies the manually produced patterning. In the future, we will exclude judgments based on one single instance of disagreement. In all these cases, the concordances were either highly metaphorical or simply odd even in a broader context.

It is to be emphasized that RG does neither immediately assess the quality of the entry nor the quality of the annotation. The output has to be analyzed by a human, since the causes of disagreements often lie in the natural ambiguity or vagueness of the concordances, whereas the categories can be perfectly distinct.

This tool is generic enough to be used with any kind of annotation where the tagset is to be optimized during the annotation.

Note

[1] This work has been using language resources developed and/or stored and/or distributed by the LINDAT-CLARIN project of the Ministry of Education of the Czech Republic (project LM2010013), and it was financed by the Czech Science Foundation under the projects P103/12/G084 and P406/2010/0875.

References

A. Dictionaries

Rundell, M. (ed.) 2007. *Macmillan English Dictionary for Advanced Learners.* (Second Edition.) Oxford: Macmillan Education. (MEDAL2).

Sinclair J. and P. Hanks et al. 1991. *Collins Cobuild English Dictionary for Advanced Learners.* HarperCollins Publishers.

B. Other literature

Aharon, R. B., I. Szpektor and I. Dagan 2010. 'Generating Entailment Rules from FrameNet.' In *Proceedings of the ACL 2010 Conference Short Papers. Presented at the ACL 2010*. Uppsala: Association for Computational Linguistics, 241–246.

Artstein, R. and M. Poesio 2008. 'Inter-coder Agreement for Computational Linguistics.' *Computational Linguistics* 34.4: 555–596.

British National Corpus, Version 3 (BNC XML Edition) (British National Corpus Consortium, 2007) http://www.natcorp.ox.ac.uk/.

Cinková, S. and P. Hanks. *Validation of Corpus Pattern Analysis – Assigning pattern numbers to random verb samples*. 15 March 2012. http://ufal.mff.cuni.cz/spr/data/publications/annotation_manual.pdf.

Cinková, S., M. Holub and V. Kríž. To appear 1. 'Managing Uncertainty in Semantic Tagging.' Accepted for *EACL 2012*, Avignon, France.

Cinková S., M. Holub, A. Rambousek and L. Smejkalová. To appear 2. 'A database of semantic clusters of verb usages.' Accepted for *LREC 2012*, Istanbul, Turkey.

Erk, K. 2010. 'What Is Word Meaning, Really? (And How Can Distributional Models Help Us Describe It?).' In R. Basili and Marco Pennacchiotti (eds.), *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*. Uppsala: Association for Computational Linguistics, 17–26.

Fellbaum, C. 1998. *WordNet. An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, C. J. and B. T. S. Atkins 1994. 'Starting Where the Dictionaries Stop: The Challenge for Computational Lexicography.' In B. T. S. Atkins and A. Zampolli (eds.), *Computational Approaches to the Lexicon*. New York: Oxford University Press, 349–393.

Hanks, P. Forthcoming. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Hanks, P. and J. Pustejovsky 2005. 'A Pattern Dictionary for Natural Language Processing', *Revue Francaise De Linguistique Appliquée* 2005.2: 63–82.

Hovy, E., M. Mitchell, M. Palmer, L. Ramshaw and R. Weischedel 2006. 'OntoNotes: The 90% Solution', in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06 . Stroudsburg, PA, USA: Association for Computational Linguistics, 57–60.

Kilgarriff, A. 1997. 'I Don't Believe in Word Senses.' *Computers and the Humanities* 31: 91–113.

Palmer, M., D. Gildea and P. Kingsbury 2005. 'The Proposition Bank: An Annotated Corpus of Semantic Roles' *Computational Linguistics Journal*, 31.1: 71–105.

Rumshisky, A., M.Verhagen and J. Moszkowicz 2009. T*he Holy Grail of Sense Definition: Creating a Sense-Disambiguated Corpus from Scratch*. Presented at the Fifth

International Workshop on Generative Approaches to the Lexicon (GL 2009), Pisa, Italy.

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson and J. Scheffczyk 2010. *FrameNet II: Extended Theory and Practice* (ICSI, University of Berkeley, 2010). 15 March, 2012. https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Weischedel, R., et al. 2011. *OntoNotes Release 4.0*. Philadelphia: Linguistic Data Consortium.

Zanzotto, F.M., M.Pennacchiotti and A.Moschitti 2009. 'A machine learning approach to textual entailment recognition.' *Natural Language Engineering* 15, 551–582.