Särtryck ur

Nordiska studier i lexikografi 11

Rapport från
Konferensen om lexikografi i Norden
Lund 24–27 maj 2011

Maintaining consistency of monolingual verb entries with interannotator agreement

Silvie Cinková, Lenka Smejkalová, Anna Vernerová,
Jonáš Thál & Martin Holub

Skrifter utgivna av Nordiska föreningen för lexikografi
Skrift nr 12

Maintaining consistency of monolingual verb entries with interannotator agreement

Silvie Cinková, Lenka Smejkalová, Anna Vernerová, Jonáš Thál & Martin Holub

We study the verb *throw* from the point of the inter-annotator agreement in the Word Sense Disambiguation (WSD) discipline. The most frequently used verbs are often polysemous. This poses a challenge when creating manually annotated data for machine learning, since polysemy threatens inter-annotator agreement. We argue that the classical WSD setup (selecting just one matching sense from a list) is inadequate for semantically complex verbs.

Keywords: Corpus Pattern Analysis, interannotator agreement, English verbs, vagueness, polysemy, word senses

1. Introduction

This study[1] approaches grouping readings (lexical units, senses) from the perspective of Natural Language Processing (NLP). A classical task in NLP is *Word Sense Disambiguation (WSD)*. Automatic WSD is mostly performed with a method of (statistical) *machine learning.* Machine learning is used in complex tasks for which humans are not able to create explicit rules. The computer learns to mimic human judgment from a *gold standard data* set with examples of

[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). It has been supported by the Czech Science Foundation projects P103/12/G084 and P406/2010/0875 as well as by the Czech Ministry of Education project MŠMT ČR LC536 and EuroMatrixPlus (FP7-IST-5-034434-IP, 7E09003). We also thank Patrick W. Hanks, Pavel Rychlý and Adam Rambousek for a comprehensive CPA guidance, the permission to experiment on a copy of PDEV, as well as for the technical assistance we have been receiving.

good solutions of the given problem. It is supposed to apply the same strategies on unknown data, so that humans perceive its decisions as acceptable.

The gold standard data for WSD usually consists of a text corpus that is interlinked with a machine-readable lexicon. Originally, digitized dictionaries for human users were used. Later on, specialized lexical resources for NLP were developed (e.g. Miller 1995 and Fellbaum (ed.) 1998, Ruppenhofer et al. 2002, Palmer et al. 2005). Yet despite all efforts put into WSD in the last two decades, the issue still remains unsolved.

The typical setup of a WSD manual annotation task is providing human annotators with a list of senses for each word. Each annotator is allowed to choose exactly one sense per word occurrence. *Interannotator agreement* (IAA) is measured several times or continuously during the task by having several annotators process the same data. Manual annotation, as well as IAA measurement, is a standard method in computational linguistics.

Semantic tasks turn out to be difficult for humans. The annotators tend to disagree with each other and thus IAA drops in semantic tasks. This is an important issue, since the computer will never make better human-like judgments than humans themselves do. If annotators in a WSD task provide each a different answer for a task, the computer has no chance to learn any strategy for assigning word senses either. Finding an adequate way to create gold standard data is therefore vital for a satisfactory WSD or any other method of semantic interpretation.

2. The different types of polysemy

WSD works with the tacit assumption that word senses are mutually exclusive; i.e. that the words are ambiguous. We will argue that the classical WSD setup is bound to fail with words that denote vague concepts and show that the relation between two intuitively distinct readings of a verb is quite often that of vagueness rather than that of ambiguity, depending on the context. The following sections will explain the notions *ambiguity* and *vagueness* (Van Deemter 2010).

2.1 Ambiguity

When a word or an expression is *ambiguous*, it denotes several concepts that are cognitively distant. When using a clearly ambiguous word such as the noun *spring*, the speaker can impossibly mean more than one of its readings at the same time *(source of water, season, coil)* or "something in between" or even "not care which the recipient chooses, all being equally adequate for conveying the

message". If in this case the recipient is unsure about the correct interpretation, it is the speaker's fault.

2.2 Vagueness

Opposite of ambiguity is *vagueness* (van Deemter 2010). A concept is *vague*, when it allows for borderline cases. For instance, there is a general understanding of the adjective *tall*, but, confronted with a row of men of different height, people will disagree on individual men whether they are tall or not. In the language, a word denoting a concept is vague when there are no clear borders between its readings and there are sensible contexts in which the speaker can mean both at the same time, or something in between, or where this information may remain underspecified without compromising the message.

Let us have the noun *glass* as an example of a vague concept. MEDO (Rundell et al. 2002) gives it three contemporary readings: hard clear substance, small container for drinking and (a summarizing term for) attractive artifacts made of glass *(a collection of Italian glass)*.

These readings have many common features: fragility, smooth surface, translucency, etc. and differ mainly in accentuating either the raw material or the (number of) artifacts made of it. In many cases, an underspecifying context can be perceived as ambiguous, since it would matter whether the speaker means the raw material or the artifact(s). Despite the three readings being intuitively easy to keep apart, many everyday contexts operate with a much coarser granularity of the semantic grouping, and this is where we can speak about vagueness. For instance, *glass polisher* is meant for glass in any form, and neither the speaker nor the recipient would care whether the material or artifacts made of it are meant. Cruse (1986) calls the relation among different readings of a word based on emphasizing one aspect of the more or less common denotate *semantic modulation*. For the purpose of this paper, we regard a concept as a *vague concept* when it is expressed by a set of mutually modulated readings of one word. The actual uses of that word in contexts where more readings can be considered can be regarded as *vague contexts.*

3. Verbs as vague concepts

Dividing verbs into readings is generally considered to be more controversial than doing the same with nouns, which we claim has to do with the inherent vagueness of most verbs. While nouns typically denote entities (event nouns disregarded), verbs typically denote events that entities undergo, as well as rela-

tions between entities. It is astonishing that our languages contain so many fewer verbs than nouns. In addition, their distribution is not even. Out of over 6,000 verb types in the English BNC corpus, the first most frequent 1,000 verbs cover more than 90% of all verb occurrences in the corpus. We apparently need only 1,000 words to describe almost all kinds of events and relations between all thinkable entities! It will be argued that this enormous descriptive power makes verbs, at least the most frequent lexical verbs, vague concepts.

If different people were to create the same verb entry independently of each other by making semantic clusters from random concordances, the number of clusters and the distribution of the concordances would be individual. Our intuitive criterion is namely the perceived similarity of the respective events, which is individual.

The similarity perception is partly affected by the morphosyntactic relations between the verb and its immediate context: e.g. any other instance of *throw* will be perceived as different from *throw up* (i.e. *vomit*), due to the particle alone. When the morphosyntactic realization of a predicate is the same, the inherent semantic characteristics of the participating entities play an important role in judging to what extent two events denoted by the same verb are similar: when *a boy throws stones*, is it the same event as when *somebody is thrown into the air and carried down the road by the motorbike* or *by an explosion*? Some similarity judgments are likely to be shared across the language community (*horse throwing a man* is different from *dawn throwing sunlight*), but some will inevitably be clustered differently by different speakers (and even differently by the same speaker at different times), depending on the granularity with which they regard the respective concordances as instances of *throwing*.

The difference in plausible contextual synonyms of *throw* between the respective clusters shows that the clusters reflect our perception of them as different sorts of events:

A boy was throwing/hurling/tossing/blowing stones.
Fast driving on gravel roads throws up rocks which can scar the car. /My wheels spit gravel and I long for a bigger windshield.
... fragmental material thrown into the air by explosive volcanic activity. and The volcano was throwing/spewing stones and lava.
Dawn threw/cast sunlight across the ruins of the old city.

To show the semantic differences between different cases of *throwing*, the contextual synonyms must have more collocational restrictions than the original verb. E.g. *spew* has more specific requirements on the agent as well as on the patient than *throw* has. The "spewer" must act as a sort of container and have a mouth or a mouth-like aperture, out of which the material is forced. Almost

anything in the world can *throw* almost anything, but only a subset of them *spew* when they *throw*.

Verbs with collocational restrictions, such as *spew*, explicitly express those semantic features of their arguments that are relevant to the event they denote. Using them, the speaker explicitly shares his world knowledge with the recipient. The use of frequent lexical verbs like *throw*, on the contrary, seems to shift the entire burden of world knowledge onto the recipient. For instance, to decode the message of the sentence *He threw the bread to the birds*, the recipient has to know that the implicit but probably the most *relevant* message concerns feeding the birds or discarding the bread rather than the explicitly mentioned propelling of bread crumbs, unlike e.g. *throwing darts*, although the hand-propelling motion is present in both. There is no need to decide which it is, because the relation between these readings is not ambiguity, but vagueness. In this way, using a semantically complex (vague) verb enables us to conflate a number of different aspects of the same event into one predicate, while the relevance of the different readings is assessed on the basis of the world knowledge (by both the speaker and the recipient).

Whenever the relations between readings in verb entries are those of vagueness rather than ambiguity, then the classical setup of the word-sense disambiguation task is completely out of place and the problem must be approached in a different way.

4. Vagueness of *throw* in semantic definitions and in the data

4.1 *throw* in OntoNotes

The inventory of wordsenses in the English OntoNotes 4.0 (Weischedel et al. 2010), an extensive linguistic resource, has arisen by merging the very granular WordNet senses (difficult for annotation) to achieve the right granularity with which two annotators reach a 90% agreement. The senses are:

1. Propel or hurl something/someone with force; toss or put forward an idea or gesture.
2. Discard, dispose, expel, or get rid of.
3. Manipulate or move something in order to operate.
4. Cast, emit, or radiate (including metaphoric expressions as *cast light, understanding*, etc.)
5. Confuse or bewilder
6. Hold an event (including *throw a fit*)
7. Form or shape, as pottery

8. Lose, as a game, intentionally
9. Give up, quit (including *throw in, throw in the towel*)
10. throw up: Vomit
11. throw in: Add as an extra or gratuity
12. None of the above

The verb *throw* has 11 senses (plus one "trash bin") now and the interannotator agreement 79%, still below the desired limit. Further merging can be expected in the next release.

Some senses are easy to match to concordances, because they are distinguished by particles (*throw in, throw up*), and they are also semantically distinct from the physical throwing as propelling an object with force or acting as the propelling force. Senses 8 and 9 are similar (both mean a sort of *giving in*), but one is realized by a particle verb or an easily identifiable idiom (*towel*), while the other combines almost exclusively with synonyms of *match, game* and *elections*. So does Sense 6 (with the typical collocate being *party*). An additional annotator convention specifies that *fit* (i.e. also *tantrum* and *wobbly*) also belongs here, since the definition alone does not imply it. Sense 3 is also easy to identify, as the list of its typical collocates is short and homogeneous: *switch, handle, engine, car, back gear.* Sense 5 is associated with the preposition *into* and a synonym of *confusion* or *disarray*, or the idiom *out/off kilter* – also quite a homogeneous group. Sense 7 (*forming pottery*) is unmistakable in real contexts, too. So is 10 (vomiting).

The official release does unfortunately not show the disagreeing annotator judgments but only the adjudicated result with an indication that two annotators disagreed. Nevertheless, three senses stand out as the most likely causes of annotator disagreements: 1, 2 and 4. They are associated with semantically very heterogeneous events, as the annotated corpus concordances show.

Sense 1 encompasses very manifold events, for instance:

The delinquents threw a brick through the school window.
He threw sixes on both die.
She threw herself enthusiastically into the project.
The earthquake threw them onto the floor.
He threw in a couple of wisecracks during her speech.
She hangs out while the boys fight and throw money around.

Sense 4 (cast, emit or radiate) is more homogeneously determined by the semantic features of the patient, which is supposed to be a liquid, vapor, light or sound.

In the example sentences of Sense 2 in the lexicon (discard, dispose, expel, or get rid of, including *throw away* and *throw off*), all occurrences of *throw* combine with the particles *away, out* or *off*, except the idiomatic *throw something into the wind*, but the particle is not prescribed as obligatory. This suggests that e.g. *throw the paper straight into the bin* can be both classified as 1 and as 2, causing a disagreement, since both the respective senses are broad and often combine.

Surprisingly, the huge OntoNotes 4.0 corpora contain only 91 sense-annotated occurrences of *throw*.[2] Of these 91 sentences, 29 double-annotated occurrences contained disagreements to be adjudicated. The most frequent adjudications are in Senses 1 (15) and 2 (9). The other readings had only marginal adjudications (Sense 4 had two. Senses 5, 6, and 10 had one each). Senses 3, 7, 8, 9 and 11 did not occur in the annotated data at all. We have therefore no idea what agreement the annotators would reach on them.

We have browsed the 24 occurrences of *throw* in OntoNotes 4.0 that were adjudicated to 1 or 2. We were not able to see whether the disagreement was caused by an annotator error or whether the lexicon entry was adjusted according to this annotation and the annotation was not revised afterwards. The patient was in most cases money, in the sense of wasting or spending it in a hope to solve a problem (4 cases). Two cases denoted discarding something (but bare *throw* was used without any particle), and the other two denoted adding to consideration. The annotators also disagreed on throwing stones etc. as missiles (2 cases). The remaining 17 occurrences were probably disagreements between 1 and 2, finally adjudicated to either 1 or 2.

We can only guess that the commonest disagreement combinations were 1, 2 and 2, 1 in Senses 1 and 2, but even so we can observe that Senses 1 and 2 are problematic for annotation, even though they are the most generic and also most frequently matched senses in the list.

Brown (Brown et al. 2010) performed IAA experiments with WordNet vs. OntoNotes sense-annotated data. They showed that IAA is harmed by the high granularity of senses rather than by the sheer number of senses (when they are not very fine-grained). Our observation does not falsify the experimental outcome; nevertheless, it suggests that most problems within the coarse-grained annotation arise in the coarsest senses anyway.

[2] However, the low number is understandable, given that a word is massively annotated only when ITA has reached 90% on a 50-sentence random sample and the latest ITA measured on *throw* was only 79% (cf. Weischedel et al. 2010:13).

4.2 *throw* according to Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) is a corpus-supported manual method of creating dictionary entries for verbs (Hanks, in press, Hanks and Pustejovsky 2005.) The lexicographer takes a random corpus sample of 250-1000 concordances and sorts them into meaningful clusters. In addition, he considers the automatic collocation analysis provided by the Sketch Engine (Kilgarriff et al., 2004). Then he describes the prototypical morphosyntactic structure of each cluster and the prototypical lexical realization of relevant collocates. This part of a dictionary reading is called *pattern*. He also makes a paraphrase of the pattern in form of a finite clause (*implicature*). Only frequent cases get a pattern of their own. Rare and unintelligible cases are marked as unclassifiable. Borderline cases are associated to patterns, when possible, but have additional markup (*exploitations*). Many marked borderline cases in a pattern result in pattern revision (mostly in the creation of an additional pattern).

We have used the *throw*-entry in Hanks' Pattern Dictionary of English Verbs (PDEV, http://corpora.fi.muni.cz/cpa/), which had a promising 1000-concordance sample annotated by the lexicographer. It turned out, however, that the sample was not finished. Therefore we finished the annotation on our own[3]. The annotation led to an addition of new patterns and to a slight revision of Hanks' original patterns. The 1000 concordances yielded about 70 patterns in total. Most patterns were naturally phraseological units as *throw the baby out with the bathwater, throw a spanner into the works,* and *throw the towel/sponge in*, as well as *throw up*.

No effort was put into making the implicatures mutually exclusive. Some of our implicatures were clearly special cases of others. E.g. a very distinct pattern of *dead bodies being thrown* would fit into the less specific *garbage discarding* pattern. The granularity of the implicatures of the respective patterns was based entirely on the data. With the instruction to associate a concordance to the most specific cluster possible, each concordance had to match exactly one implicature – or none.

The original pattern inventory also contained a pattern for propelling physical objects towards a target and a pattern for discarding physical objects (spending assets and aiming gestures/words etc. had already been separated). The (single) annotation of the 1000 concordances uncovered a fuzzy border between propelling and discarding, which is likely to have caused disagreements

[3] A sample of 30 PDEV verbs was IAA-tested and revised. It is publicly available as VPS-30-En at http://hdl.handle.net/11858/00-097C-0000-0005-BF95-B in the LINDAT-CLARIN repository.

in the OntoNotes 4.0 annotation. Nevertheless, one can still think of many contexts in which the discarding implicature is clearly the dominant one. Therefore the self-standing *discard*-pattern for physical objects was kept and just extended with more lexical suggestions (*in the bin, in the garbage*, etc.) to encourage annotators to rely on their world knowledge, and the *target* condition was removed from the *propel* pattern. Another pattern was created for throwing physical objects at a target as *missiles*. Restructuring the patterns this way brought more homogeneous groups of concordances.

4.3 A case study of *throw* in a vague context

However, minute pattern splitting is a burden in WSD tasks, as Brown (Brown et al. 2010) has demonstrated. Any new batch of corpus data to analyze is bound to throw up concordances that do not quite match any current pattern, which in turn is bound to threaten the IAA.

The following example will show that, no matter the granularity of dictionary readings, frequent lexical verbs tend to denote several aspects of an event at once. Finding and encoding their typical combinations is very difficult and probably domain-, culture- and application dependent (cf. Kilgarriff 1997), such that petrifying them into hard-wired mutually exclusive readings would be counter-productive.

Let us have a look at a BNC sentence that describes an event by conflating several world-knowledge-based implicatures into the verb *throw*:

Osbern has his father killed by a lowly mob and thrown *to birds and wild animals.*

Five of our revised patterns are relevant for this particular event:

1. Human uses *hands* to *propel* a physical *object* in a direction for a short distance
   Ex.: Tourists are encouraged to throw *coins into the fountains for good fortune.*
2. Human violently *pushes* or shoves or kicks another *human* so that the other human loses control over his movements and *falls*
   Ex.: He threw her to the ground/against the cupboard…
3. Human *discards* or gets rid of an artifact or *stuff*
   Ex.: He threw the paper straight into the bin/threw it away, threw it out.
4. Human (*murderer*) disposes of or *discards* or *hides* the *body* of his *victim* to some place
   Ex.: Their corpses were thrown *down a well.*
5. Human *feeds an animal/animals* with a physical object or a substance
   Ex.: It was like throwing *a piece of meat to sharks.*

Now, if the annotator had to pick one reading only, she would have to decide which throwing is the most relevant one. Interannotator disagreement is inevitable here, as everyone can interpret the nuances of a text differently.

Discussion

Liberman (2009: 2) claims that:

> […] linguistic annotation tasks are not really classification problems, but rather translation problems. We don't normally assume that there is only one correct translation of a Chinese sentence into English; nor do we try to make this true by constructing elaborate translation guidelines to cover every relevant contingency, though in principle we could.

Recent experiments (Rumshisky et al. 2009, Erk 2010) as well as empirical observations of experienced lexicographers (Krishnamurthy and Nicholls 2000, Hampton, 2007) suggest that WSD in its classical setup is extremely difficult for humans. On the other hand, Rumshisky showed in a 2009 experiment with the Mechanical Turk (Rumshisky et al. 2009) that even linguistic non-experts reach a good agreement on deciding whether two occurrences of a word (verb) are used in the same sense. In addition, their groupings turned out to be quite similar to a sense grouping made by a professional lexicographer, only coarser. Erk (Erk 2010) reported a similar result with concordances compared to WordNet (Miller 1995 and Fellbaum 1998) senses. Wilks and Ide (Wilks and Ide 2006) question the usefulness of WSD being performed as a separate task in NLP applications in general.

Conclusion

Experiments with the traditional as well as the more recent approaches to human WSD annotation prove that humans are unable to pick "the right" sense from a predefined list to decode a word in context. This has two implications:

1) "picking the right sense" from a lexicon is not exactly the way in which humans use dictionaries to look up an unfamiliar word, although we clearly make use of dictionaries when decoding texts with unfamiliar words;
2) if the computer is supposed to mimic the human way of using dictionaries for text analysis, the task of word sense disambiguation (or text understanding in general, cf. Wilks and Ide 2007) has to be constituted in a different way.

References

Brown, Susan et al., 2010: Number or Nuance: Which Factors Restrict Reliable Word Sense Annotation? In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta.

Cruse, D.A., 1986: Lexical Semantics. Cambridge University Press.

Davies, Mark, 2008: The Corpus of Contemporary American English: 425 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Van Deemter, Kees, 2010: Not exactly. In praise of vagueness. Oxford University Press.

Erk, Katrin, 2010: What Is Word Meaning, Really? (And How Can Distributional Models Help Us Describe It?) In: Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics, ACL 2010, pp. 17–26, Uppsala, Sweden

Fellbaum, Christiane (ed.), 1998: WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Hampton, J. A., 2007: Typicality, graded membership, and vagueness. Cognitive Science, 31, pp. 355–384.

Hanks, Patrick Wyndham. in press. Lexical Analysis. Norms and Exploitations. MIT Press.

Hanks, Patrick, Pustejovsky, James, 2005: A Pattern Dictionary for Natural Language Processing. In: *Revue française de linguistique appliquée* 10 (2).

Kilgarriff, Adam, Rychly, Pavel, Smrz, Pavel, Tugwell, David, 2004: The Sketch Engine. In *EURALEX 2004, Lorient, France*; Pp 105-116. http://www.sketchengine.co.uk

Kilgarriff, Adam, 1997: "I Don't Believe in Word Senses". Computers and the Humanities, 13 (2).

Krishnamurthy, R., Nicholls, D., 2000: Peeling an onion: the lexicographers' experience of manual sense-tagging. Computers and the Humanities, 34(1–2).

Liberman, Mark, 2009: The Annotation Conundrum. In: Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? Athens, Greece.

Miller, George A., 1995: WordNet: A Lexical Database for English. Communications of the ACM Vol. 38:11, pp. 39–41.

Palmer, M., Kingsbury, P., Gildea, D., 2005: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics 31:1, pp. 71–106.

Rumshisky, Anna et al., 2009: The Holy Grail of Sense Definition: Creating a Sense-Disambiguated Corpus from Scratch. In: Fifth International Workshop on Generative Approaches to the Lexicon (GL 2009). Pisa, Italy.

Rundell, M. et al., 2002: Macmillan English Dictionary for Advanced Learners. International Student Edition. MacMillan Education. Oxford. (online access: http://online.macmillandictionary. com/mc_au2/macmil.htm)

Ruppenhofer, Josef, Collin F. Baker and Charles J. Fillmore, 2002: The FrameNet Database and Software Tools. In: Braasch, Anna and Claus Povlsen (eds.), Proceedings of the Tenth Euralex International Congress. Copenhagen, Denmark. Vol. I, pp. 371–375.

Weischedel, A. et al., 2010: OntoNotes Release 4.0. Linguistic Data Consortium, Philadelphia

Wilks, Yorick, Ide, Nancy, 2007: Making Sense about Sense. Word Sense Disambiguation, Algorithms and Applications (pp. 47–73). Springer.

Silvie Cinková
Charles University in Prague, Faculty of Mathematics and Physics - Institute of Formal and Applied Linguistics, Malostranské nám. 25, 118 00 Praha 1, Czech Republic.
cinkova@ufal.mff.cuni.cz

Lenka Smejkalová
Charles University in Prague.
smejkalova@ufal.mff.cuni.cz

Anna Vernerová
Charles University in Prague.
vernerova@ufal.mff.cuni.cz

Jonáš Thál
jonlatash@gmail.com

Martin Holub
Charles University in Prague.
holub@ufal.mff.cuni.cz