Managing Uncertainty in Semantic Tagging

Silvie Cinková and **Martin Holub** and **Vincent Kríž**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{cinkova|holub}@ufal.mff.cuni.cz
vincent.kriz@gmail.com

Abstract

Low interannotator agreement (IAA) is a well-known issue in manual semantic tagging (sense tagging). IAA correlates with the granularity of word senses and they both correlate with the amount of information they give as well as with its reliability. We compare different approaches to semantic tagging in WordNet, FrameNet, PropBank and OntoNotes with a small tagged data sample based on the Corpus Pattern Analysis to present the *reliable information gain* (RG), a measure used to optimize the semantic granularity of a sense inventory with respect to its reliability indicated by the IAA in the given data set. RG can also be used as feedback for lexicographers, and as a supporting component of automatic semantic classifiers, especially when dealing with a very fine-grained set of semantic categories.

1 Introduction

The term *semantic tagging* is used in two divergent areas:

1) recognizing objects of semantic importance, such as entities, events and polarity, often tailored to a restricted domain, or

2) relating occurrences of words in a corpus to a lexicon and selecting the most appropriate semantic categories (such as *synsets, semantic frames, wordsenses, semantic patterns* or *framesets*).

We are concerned with the second case, which seeks to make lexical semantics tractable for computers. Lexical semantics, as opposed to propositional semantics, focuses the meaning of lexical items. The disciplines that focus lexical semantics are lexicology and lexicography rather than logic. By semantic tagging we mean a process of assigning semantic categories to target words in given contexts. This process can be either manual or automatic.

Traditionally, semantic tagging relies on the tacit assumption that various uses of polysemous words can be sorted into discrete senses; understanding or using an unfamiliar word be then like looking it up in a dictionary. When building a dictionary entry for a given word, the lexicographer sorts a number of its occurrences into discrete *senses* present (or emerging) in his/her mental lexicon, which is supposed to be shared by all speakers of the same language. The assumed common mental representation of a words meaning should make it easy for other humans to assign random occurrences of the word to one of the pre-defined senses (Fellbaum et al., 1997).

This assumption seems to be falsified by the interannotator agreement (IAA, sometimes ITA) constantly reported much lower in semantic than in morphological or syntactic annotation, as well as by the general divergence of opinion on which value of which IAA measure indicates a reliable annotation. In some projects (e.g. OntoNotes (Hovy et al., 2006)), the percentage of agreements between two annotators is used, but a number of more complex measures are available (for a comprehensive survey see (Artstein and Poesio, 2008)). Consequently, using different measures for IAA makes the reported IAA values incomparable across different projects.

Even skilled lexicographers have trouble selecting one discrete sense for a concordance (Krishnamurthy and Nicholls, 2000), and, more to say, when the tagging performance of lexicographers and ordinary annotators (students) was

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 840–850,
Avignon, France, April 23 - 27 2012.

compared, the experiment showed that the mental representations of a word's semantics differ for each group (Fellbaum et al., 1997), and cf. (Jorgensen, 1990). Lexicographers are trained in considering subtle differences among various uses of a word, which ordinary language users do not reflect. Identifying a semantic difference between uses of a word and deciding whether a difference is important enough to constitute a separate sense means presenting a word with a certain degree of *semantic granularity*. Intuitively, the finer the granularity of a word entry is, the more opportunities for interannotator disagreement there are and the lower IAA can be expected. Brown et al. proved this hypothesis experimentally (Brown et al., 2010). Also, the annotators are less confident in their decisions, when they have many options to choose from (Fellbaum et al. (1998) reported a drop in subjective annotators confidence in words with 8+ senses).

Despite all the known issues in semantic tagging, the major lexical resources (WordNet (Fellbaum, 1998), FrameNet (Ruppenhofer et al., 2010), PropBank (Palmer et al., 2005) and the word-sense part of OntoNotes (Weischedel et al., 2011)) are still maintained and their annotation schemes are adopted for creating new manually annotated data (e.g. MASC, the Manually Annotated Subcorpus (Ide et al., 2008)). More to say, these resources are not only used in WSD and semantic labeling, but also in research directions that in their turn do not rely on the idea of an inventory of discrete senses any more, e.g. in *distributional semantics* (Erk, 2010) and *recognizing textual entailment* (e.g. (Zanzotto et al., 2009) and (Aharon et al., 2010)).

It is a remarkable fact that, to the best of our knowledge, there is no measure that would relate granularity, reliability of the annotation (derived from IAA) and the resulting information gain. Therefore it is impossible to say where the optimum for granularity and IAA lies.

2 Approaches to semantic tagging

2.1 Semantic tagging vs. morphological or syntactic analysis

Manual semantic tagging is in many respects similar to morphological tagging and syntactic analysis: human annotators are trained to sort certain elements occurring in a running text according to a reference source. There is, nevertheless, a substantial difference: whereas morphologically or syntactically annotated data exist separately from the reference (tagset, annotation guide, annotation scheme), a semantically tagged resource can be regarded both as a corpus of texts disambiguated according to an attached inventory of semantic categories and as a lexicon with links to example concordances for each semantic category. So, in semantically tagged resources, the data and the reference are intertwined. Such double-faced semantic resources have also been called *semantic concordances* (Miller et al., 1993a). For instance, one of the earlier versions of WordNet, the largest lexical resource for English, was used in the semantic concordance SemCor (Miller et al., 1993b). More recent lexical resources have been built as semantic concordances from the very beginning (PropBank (Palmer et al., 2005), OntoNotes word senses (Weischedel et al., 2011)).

In morphological or syntactic annotation, the tagset or inventory of constituents are given beforehand and are supposed to hold for all tokens/sentences contained in the corpus. Problematic and theory-dependent issues are few and mostly well-known in advance. Therefore they can be reflected by a few additional conventions in the annotation manual (e.g. where to draw the line between particles and prepositions or between adjectives and verbs in past participles (Santorini, 1990) or where to attach a prepositional phrase following a noun phrase and how to treat specific "financialspeak" structures (Bies et al., 1995)). Even in difficult cases, there are hardly more than two options of interpretation. Data manually annotated for morphology or surface syntax are reliable enough to train syntactic parsers with an accuracy above 80 % (e.g. (Zhang and Clark, 2011; McDonald et al., 2006)).

On the other hand, semantic tagging actually employs a different tagset for each word lemma. Even within the same part of speech, individual words require individual descriptions. Possible similarities among them come into relief ex post rather than that they could be imposed on the lexicographers from the beginning. When assigning senses to concordances, the annotator often has to select among more than two relevant options. These two aspects make achieving good IAA much harder than in morphology and syn-

tax tasks. In addition, while a linguistically educated annotator can have roughly the same idea of parts of speech as the author of the tagset, there is no chance that two humans (not even two professional lexicographers) would create identical entries for e.g. a polysemous verb. Any human evaluation of complete entries would be subjective. The maximum to be achieved is that the entry reflects the corpus data in a reasonable granular way on which annotators still can reach reasonable IAA.

2.2 Major existing semantic resources

The granularity vs. IAA equilibrium is of great concern in creating lexical resources as well as in applications dealing with semantic tasks. When WordNet (Fellbaum, 1998) was created, both IAA and subjective confidence measurements served as an informal feedback to lexicographers (Fellbaum et al., (1998), p. 200). In general, WordNet has been considered a resource too fine-grained for most annotations (and applications). Navigli (2006) developed a method of reducing the granularity of WordNet by mapping the synsets to senses in a more coarse-grained dictionary. A manual, more coarse-grained grouping of WordNet senses has been performed in OntoNotes (Weischedel et al., 2011). The OntoNotes 90 % solution (Hovy et al., 2006) actually means such a degree of granularity that enables a 90-%-IAA. OntoNotes is a reaction to the traditionally poor IAA in WordNet annotated corpora, caused by the high granularity of senses. The quality of semantic concordances is maintained by numerous iterations between lexicographers and annotators. The categories 'right'–'wrong' have been, for the purpose of the annotated linguistic resource, defined by the IAA score, which is—in OntoNotes—calculated as the percentage of agreements between two annotators.

Two other, somewhat different, lexical resources have to be mentioned to complete the picture: FrameNet (Ruppenhofer et al., 2010) and PropBank (Palmer et al., 2005). While WordNet and OntoNotes pair words and word senses in a way comparable to printed lexicons, FrameNet is primarily an inventory of *semantic frames* and PropBank focuses the *argument structure* of verbs and nouns (NomBank (Meyers et al., 2008), a related project capturing the argument structure of nouns, was later integrated in OntoNotes).

In FrameNet corpora, content words are associated to particular semantic frames that they evoke (e.g. *charm* would relate to the *Aesthetics* frame) and their collocates in relevant syntactic positions (arguments of verbs, head nouns of adjectives, etc.) would be assigned the corresponding *frame-element* labels (e.g. in *their dazzling charm*, *their* would be The Entity for which a particular gradable Attribute is appropriate and under consideration and *dazzling* would be Degree). Neither IAA nor granularity seem to be an issue in FrameNet. We have not succeeded in finding a report on IAA in the original FrameNet annotation, except one measurement in progress in the annotation of the Manually Annotated Subcorpus of English (Ide et al., 2008).[1]

PropBank is a valency (argument structure) lexicon. The current resource lists and labels arguments and obligatory modifiers typical of each (very coarse) word sense (called *frameset*). Two core criteria for distinguishing among framesets are the semantic roles of the arguments along with the syntactic alternations that the verb can undergo with that particular argument set. To keep low granularity, this lexicon—among other things—does usually not make special framesets for metaphoric uses. The overall IAA measured on verbs was 94 % (Palmer et al., 2005).

2.3 Semantic Pattern Recognition

From corpus-based lexicography to semantic patterns

The modern, corpus-based lexicology of 1990s (Sinclair, 1991; Fillmore and Atkins, 1994) has had a great impact on lexicography. There is a general consensus that dictionary definitions need to be supported by corpus examples. Cf. Fellbaum (2001):

"For polysemous words, dictionaries [...] do not say enough about the range of possible contexts that differentiate the senses. [...] On the other hand, texts or corpora [...] are not explicit about the word's meaning. When we first encounter a new word in a text, we can usually form only a vague idea of its meaning; checking a dictionary will clarify the meaning. But the more contexts we encounter for a word, the harder it is to match them against only one dictionary sense."

[1] Checked on the project web www.anc.org/MASC/Home 2011-10-29.

The lexical description in modern English monolingual dictionaries (Sinclair et al., 1987; Rundell, 2002) 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 the modern corpus-based lexicography: in classical dictionaries as well as in semantic concordances.

Nevertheless, the word senses, even when obtained by a collective mind of lexicographers and annotators, are naturally hard-wired and tailored to the annotated corpus. They may be too fine-grained or too coarse-grained for automatic processing of different corpora (e.g. a restricted-domain corpus). Kilgarriff (1997, p. 115) shows (the *handbag* example) that *there is no reason to expect the same set of word senses to be relevant for different tasks* and that *the corpus dictates the word senses* and therefore 'word sense' *was not found to be sufficiently well-defined to be a workable basic unit of meaning* (p. 116). On the other hand, even non-experts seem to agree reasonably well when judging the similarity of use of a word in different contexts (Rumshisky et al., 2009). Erk et al. (2009) showed promising annotation results with a scheme that allowed the annotators graded judgments of similarity between two words or between a word and its definition.

Verbs are the most challenging part of speech. We see two major causes: *vagueness* and *coercion*. We neglect ambiguity, since it has proved to be rare in our experience.

CPA and PDEV

Our current work focuses on English verbs. It has been inspired by the manual Corpus Pattern Analysis method (CPA) (Hanks, forthcoming) and its implementation, the Pattern Dictionary of English Verbs (PDEV) (Hanks and Pustejovsky, 2005). PDEV is a semantic concordance built on yet a different principle than FrameNet, WordNet, PropBank or OntoNotes. The manually extracted *patterns* of frequent and normal verb uses are, roughly speaking, intuitively similar uses of a verb that express—in a syntactically similar form—a similar event in which similar participants (e.g. humans, artifacts, institutions, other events) are involved. Two patterns can be semantically so tightly related that they could appear together under one sense in a traditional dictionary. The patterns are not senses but syntactico-semantically characterized prototypes (see the example verb *submit* in Table 1). Concordances that match these prototypes well are called *norms* in Hanks (forthcoming). Concordances that match with a reservation (metaphorical uses, argument mismatch, etc.) are called *exploitations*. The PDEV corpus annotation indicates the norm-exploitation status for each concordance.

Compared to other semantic concordances, the granularity of PDEV is high and thus discouraging in terms of expected IAA. However, selecting among patterns does not really mean disambiguating a concordance but rather determining to which pattern it is most similar—a task easier for humans than WSD is. This principle seems particularly promising for verbs as words expressing events, which resist the traditional word sense disambiguation the most.

A novel approach to semantic tagging

We present the *semantic pattern recognition* as a novel approach to semantic tagging, which is different from the traditional word-sense assignment tasks. We adopt the central idea of CPA that words do not have fixed senses but that regular patterns can be identified in the corpus that activate different conversational implicatures from the meaning potential of the given verb. Our method draws on a hard-wired, fine-grained inventory of semantic categories manually extracted from corpus data. This inventory represents the maximum semantic granularity that humans are able to recognize in normal and frequent uses of a verb in a balanced corpus. We thoroughly analyze the interannotator agreement to find out which of the highly semantic categories are useful in the sense of information gain. Our goal is a dynamic optimization of semantic granularity with respect to given data and target application.

Like Passonneau et al. (2010), we are convinced that IAA is specific to each respective word and reflects its inherent semantic properties as well as the specificity of contexts the given word occurs in, even within the same balanced corpus. We accept as a matter of fact that interannotator confusion is inevitable in semantic tagging. However, the amount of uncertainty of the

No.	Pattern / Implicature
1	**[[Human 1 | Institution 1] ^ [Human 1 | Institution 1 = Competitor]] submit [[Plan | Document | Speech Act | Proposition | {complaint | demand | request | claim | application | proposal | report | resignation | information | plea | petition | memorandum | budget | amendment | programme | ...}] ^ [Artifact | Artwork | Service | Activity | {design | tender | bid | entry | dance | ...}]] (({to} Human 2 | Institution 2 = authority)^({to} Human 2 | Institution 2 = referee)) ({for} {approval | discussion | arbitration | inspection | designation | assessment | funding | taxation | ...})** [[Human 1 | Institution 1]] presents [[Plan | Document]] to [[Human 2 | Institution 2]] for {approval | discussion | arbitration | inspection | designation | assessment | taxation | ...}
2	**[Human | Institution] submit [THAT-CL|QUOTE]** [[Human | Institution]] respectfully expresses {that [CLAUSE]} and invites listeners or readers to accept that {that [CLAUSE]} is true}
4	**[Human 1 | Institution 1] submit (Self) ({to} Human 2 | Institution 2)** [[Human 1 | Institution 1]] acknowledges the superior force of [[Human 2 | Institution 2]] and puts [[Self]] in the power of [[Human 2 | Institution 2]]
5	**[Human 1] submit (Self) [[{to} Eventuality = Unpleasant] ^ [{to} Rule]]** [[Human 1]] accepts [[Rule |Eventuality = Unpleasant]] without complaining
6	**[passive]** **[Human| Institution] submit [Anything] [{to} Eventuality]** [[Human 1|Institution 1]] exposes [[Anything]] to [[Eventuality]]

Table 1: Example of patterns defined for the verb *submit*.

"right" tag differs a lot, and should be quantified. For that purpose we developed the *reliable information gain* measure presented in Section 3.2.

CPA Verb Validation Sample

The original PDEV had never been tested with respect to 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 inconsistencies. We put down an annotation manual (a momentary snapshot of the theory) and trained three annotators accordingly. For practical annotation we use the infrastructure developed at Masaryk University in Brno (Horák et al., 2008), which was also used for the original PDEV development. After initial IAA experiments with the original PDEV, we decided to select 30 verb entries from PDEV along with the annotated concordances. We made a new semantic concordance sample (Cinková et al., 2012) for the validation of the annotation scheme. We refer to this new collection[2] as VPS-30-En (Verb Pattern Sample, 30 English verbs).

We slightly 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 and a test sample of 50 concordances for annotation. We measured IAA, using Fleiss's kappa,[3] and analyzed the interannotator confusion manually. IAA varied from verb to verb, mostly reaching safely above 0.6. When the IAA was low and the type of confusion indicated a problem in the entry, the entry was revised. 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 went through one more additional procedure, the *adjudication*: first, the lexicographer compared the three annotations and eliminated evident errors. Then the lexicographer selected one value for each concordance to remain in the resulting one-value-per-concordance gold standard data and recorded it into the gold standard set. The adjudication pro-

[2]This new lexical resource, including the complete documentation, is publicly available at http://ufal.mff.cuni.cz/spr.

[3]Fleiss's kappa (Fleiss, 1971) is a generalization of Scott's π statistic (Scott, 1955). In contrast to Cohen's kappa (Cohen, 1960), Fleiss's kappa evaluates agreement between multiple raters. However, Fleiss's kappa is *not* a generalization of Cohen's kappa, which is a different, yet related, statistical measure. Sometimes, the terminology about kappas is confusing in the literature. For a detailed explanation refer e.g. to (Artstein and Poesio, 2008).

tocol has been kept for further experiments. All values except the marked errors are regarded as equally acceptable for this type of experiments. In the end, we get for each verb:

- an entry, which is an inventory of semantic categories (patterns)
- 300+ manually annotated concordances (single values)
- out of which 50 are manually annotated and adjudicated concordances (multiple values without evident errors).

3 Tagging confusion analysis

3.1 Formal model of tagging confusion

To formally describe the semantic tagging task, we assume a target word and a (randomly selected) corpus sample of its occurrences. The tagged sample is $\mathcal{S} = \{s_1, \ldots, s_r\}$, where each *instance* s_i is an occurrence of the target word with its context, and r is the sample size.

For multiple annotation we need a set of m annotators $\mathcal{A} = \{A_1, \ldots, A_m\}$ who choose from a given set of semantic categories represented by a set of n semantic tags $\mathcal{T} = \{t_1, \ldots, t_n\}$. Generally, if we admitted assigning more tags to one word occurrence, annotators could assign any subset of $\mathcal{T}$ to an instance. In our experiments, however, annotators were allowed to assign just one tag to each tagged instance. Therefore each annotator is described as a function that assigns a single member set to each instance $A_i(s) = \{t\}$, where $s \in \mathcal{S}$, $t \in \mathcal{T}$. When a pair of annotators tag an instance s, they produce a set of one or two different tags $\{t, t'\} = A_i(s) \cup A_j(s)$.

Detailed information about interannotator (dis)agreement on a given sample $\mathcal{S}$ is represented by a set of $\binom{m}{2}$ symmetric matrices $C_{ij}^{A_k A_l} = |\{s \in \mathcal{S} \mid A_k(s) \cup A_l(s) = \{t_i, t_j\}\}|$, for $1 \leq k < l \leq m$, and $i, j \in \{1, \ldots, n\}$. Note that each of those matrices can be easily computed as $C^{A_k A_l} = C + C^T - I_n C$, where C is a conventional confusion matrix representing the agreement between annotators A_k and A_l, and I_n is a unit matrix.

Definition: *Aggregated Confusion Matrix* (ACM)

$$C^\star = \sum_{1 \leq k < l \leq m} C^{A_k A_l}.$$

Properties: ACM is symmetric and for any $i \neq j$ the number $C_{ij}^\star$ says how many times a pair of annotators disagreed on two tags t_i and t_j, while $C_{ii}^\star$ is the frequency of agreements on t_i; the sum in the i-th row $\sum_j C_{ij}^\star$ is the total frequency of assigned sets $\{t, t'\}$ that contain t_i.

An example of ACM is given in Table 2. The corresponding confusion matrices are shown in Table 3.

	1	1.a	2	4	5
1	85	8	2	0	0
1.a	8	1	2	0	0
2	2	2	34	0	0
4	0	0	0	4	8
5	0	0	0	8	6

Table 2: Aggregated Confusion Matrix.

Our approach to exact tagging confusion analysis is based on probability and information theory. Assigning semantic tags by annotators is viewed as a random process. We define (categorical) random variable T_1 as the outcome of one annotator; its values are single member sets $\{t\}$, and we have mr observations to compute their probabilities. The probability that an annotator will use t_i is denoted by $p_1(t_i) = \Pr(T_1 = \{t_i\})$ and is practically computed as the relative frequency of t_i among all mr assigned tags. Formally,

$$p_1(t_i) = \frac{1}{mr} \sum_{k=1}^{m} \sum_{j=1}^{r} |A_k(s_j) \cap \{t_i\}|.$$

The outcome of two annotators (they both tag the same instance) is described by random variable T_2; its values are single or double member sets $\{t, t'\}$, and we have $\binom{m}{2} r$ observations to compute their probabilities. In contrast to p_1, the probability that t_i will be used by a *pair* of annotators is denoted by $p_2(t_i) = \Pr(T_2 \supseteq \{t_i\})$, and is computed as the relative frequency of assigned sets $\{t, t'\}$ containing t_i among all $\binom{m}{2} r$ observations:

$$p_2(t_i) = \frac{1}{\binom{m}{2} r} \sum_k C_{ik}^\star.$$

We also need the conditional probability that an annotator will use t_i given that *another* annotator has used t_j. For convenience, we use the notation $p_2(t_i \mid t_j) = \Pr(T_2 \supseteq \{t_i\} \mid T_2 \supseteq \{t_j\})$.

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

Table 3: Example of all confusion matrices for the target word *submit* and three annotators.

Obviously, it can be computed as

$$p_2(t_i \mid t_j) = \frac{\Pr(T_2 = \{t_i, t_j\})}{\Pr(T_2 \supseteq \{t_j\})} = \frac{C^\star_{ij}}{\binom{m}{2} r \cdot p_2(t_j)} = \frac{C^\star_{ij}}{\sum_k C^\star_{jk}}.$$

Definition: *Confusion Probability Matrix* (CPM)

$$C^p_{ji} = p_2(t_i \mid t_j) = \frac{C^\star_{ij}}{\sum_k C^\star_{jk}}.$$

Properties: The sum in any row is 1. The j-th row of CPM contains probabilities of assigning t_i given that *another* annotator has chosen t_j for the same instance. Thus, the j-th row of CPM describes *expected tagging confusion* related to the tag t_j.

An example is given in Table 3 (all confusion matrices for three annotators), in Table 2 (the corresponding ACM), and in Table 4 (the corresponding CPM).

	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

Table 4: Example of Confusion Probability Matrix.

3.2 Semantic granularity optimization

Now, having a detailed analysis of expected tagging confusion described in CPM, we are able to compare usefulness of different semantic tags using a measure of the information content associated with them (in the information theory sense). Traditionally, the amount of self-information contained in a tag (as a probabilistic event) depends only on the probability of that tag, and would be defined as $I(t_j) = -\log p_1(t_j)$. However, intuitively one can say that a good measure of usefulness of a particular tag should also take into consideration the expected tagging confusion related to the tag. Therefore, to exactly measure usefulness of the tag t_j we propose to compare and measure similarity of the distribution $p_1(t_i)$ and the distribution $p_2(t_i \mid t_j)$, $i = 1, \ldots, n$. How much information do we gain when an annotator assigns the tag t_j to an instance? When the tag t_j has once been assigned to an instance by an annotator, one would naturally expect that *another* annotator will *probably tend* to assign the same tag t_j to the same instance. Formally, things make good sense if $p_2(t_j \mid t_j) > p_1(t_j)$ and if $p_2(t_i \mid t_j) < p_1(t_i)$ for any i different from j. If $p_2(t_j \mid t_j) = 100\,\%$, then there is full consensus about assigning t_j among annotators; then and only then the measure of usefulness of the tag t_j should be maximal and should have the value of $-\log p_1(t_j)$. Otherwise, the value of usefulness should be smaller. This is our motivation to define a quantity of *reliable* information gain obtained from semantic tags as follows:

Definition: *Reliable Gain* (RG) from the tag t_j is

$$RG(t_j) = \sum_k -(-1)^{\delta_{kj}} p_2(t_k|t_j) \log \frac{p_2(t_k|t_j)}{p_1(t_k)}.$$

Properties: RG is similar to the well known Kullback-Leibler divergence (or information gain). If $p_2(t_i \mid t_j) = p_1(t_i)$ for all $i = 1, \ldots, n$, then $RG(t_j) = 0$. If $p_2(t_j \mid t_j) = 100\,\%$, then and only then $RG(t_j) = -\log p_1(t_j)$, which is the maximum. If $p_2(t_i \mid t_j) < p_1(t_i)$ for all i different from j, the greater difference in probabilities, the bigger (and positive) $RG(t_j)$. And vice versa, the inequality $p_2(t_i \mid t_j) > p_1(t_i)$ for all i different from j implies a negative value of $RG(t_j)$.

Definition: *Average Reliable Gain* (ARG) from the tagset $\{t_1, \ldots, t_n\}$ is computed as an expected value of $RG(t_j)$:

$$ARG = \sum_j p_1(t_j) RG(t_j)$$

Properties: ARG has its maximum value if the CPM is a unit matrix, which is the case of the absolute agreement among all annotators. Then ARG has the value of the entropy of the p_1 distribution: $ARG_{max} = H(p_1(t_1), \ldots, p_1(t_n))$.

Merging tags with poor RG

The main motivation for developing the ARG value was the optimization of the tagset granularity. We use a semi-greedy algorithm that searches for an "optimal" tagset. The optimization process starts with the fine-grained list of CPA semantic categories and then the algorithm merges some tags in order to maximize the ARG value. An example is given in Table 5. Tables 6 and 7 show the ACM and the CPM after merging. The examples relate to the verb *submit* already shown in Tables 1, 2, 3 and 4.

Original tagset			Optimal merge		
Tag	f	RG	Tag	f	RG
1	90	+0.300	1 + 1.a	96	+0.425
1.a	6	−0.001			
2	36	+0.447	2	36	+0.473
4	8	−0.071	4 + 5	18	+0.367
5	10	−0.054			

Table 5: Frequency and Reliable Gain of tags.

	1	2	4
1	94	4	0
2	4	34	0
4	0	0	18

Table 6: Aggregated Confusion Matrix after merging.

	1	2	4
1	0.959	0.041	0.000
2	0.105	0.895	0.000
4	0.000	0.000	1.000

Table 7: Confusion Probability Matrix after merging.

3.3 Classifier evaluation with respect to expected tagging confusion

An automatic classifier is considered to be a function c that—the same way as annotators— assigns tags to instances $s \in \mathcal{S}$, so that $c(s) = \{t\}$, $t \in \mathcal{T}$. The traditional way to evaluate the accuracy of an automatic classifier means to compare its output with the correct semantic tags on a *Gold Standard* (GS) dataset. Within our formal framework, we can imagine that we have a "gold" annotator A_g, so that the GS dataset is represented by $A_g(s_1), \ldots, A_g(s_r)$. Then the classic accuracy score can be computed as $\frac{1}{r}\sum_{i=1}^{r} |A_g(s_i) \cap c(s_i)|$. However, that approach does not take into consideration the fact that some semantic tags are quite confusing even for human annotators. In our opinion, automatic classifier should not be penalized for mistakes that would be made even by humans. So we propose a more complex evaluation score using the knowledge of the expected tagging confusion stored in CPM.

Definition: Classifier evaluation *Score* with respect to tagging confusion is defined as the proportion $Score(c) = S(c)/S_{max}$, where

$$\begin{aligned} S(c) &= \frac{\alpha}{r}\sum_{i=1}^{r} |A_g(s_i) \cap c(s_i)| \; + \\ &\quad + \frac{1-\alpha}{r}\sum_{i=1}^{r} p_2(c(s_i) \mid A_g(s_i)) \\ S_{max} &= \alpha + \frac{1-\alpha}{r}\sum_{i=1}^{r} p_2(A_g(s_i) \mid A_g(s_i)). \end{aligned}$$

	$\alpha = 1$		$\alpha = 0.5$		$\alpha = 0$	
Verb		Score		Score		Score
halt	1	0.84	2	0.90	4	0.81
submit	2	0.83	1	0.90	1	0.84
ally	3	0.82	3	0.89	5	0.76
cry	4	0.79	4	0.88	2	0.82
arrive	5	0.74	5	0.85	3	0.81
plough	6	0.70	6	0.81	6	0.72
deny	7	0.62	7	0.74	7	0.66
cool	8	0.58	8	0.69	8	0.53
yield	9	0.55	9	0.67	9	0.52

Table 8: Evaluation with different α values.

Table 8 gives an illustration of the fact that using different α values one can get different re-

sults when comparing tagging accuracy for different words (a classifier based on bag-of-words approach was used). The same holds true for comparison of different classifiers.

3.4 Related work

In their extensive survey article Artstein and Poesio (2008) state that word sense tagging is one of the hardest annotation tasks. They assume that making distinctions between semantic categories must rely on a dictionary. The problem is that annotators often cannot consistently make the fine-grained distinctions proposed by trained lexicographers, which is particularly serious for verbs, because verbs generally tend to be polysemous rather than homonymous.

A few approaches have been suggested in the literature that address the problem of the fine-grained semantic distinctions by (automatic) measuring sense distinguishability. Diab (2004) computes sense perplexity using the entropy function as a characteristic of training data. She also compares the sense distributions to obtain sense distributional correlation, which can serve as a "very good direct indicator of performance ratio", especially together with sense context confusability (another indicator observed in the training data). Resnik and Yarowsky (1999) introduced the communicative/semantic distance between the predicted sense and the "correct" sense. Then they use it for evaluation metric that provides partial credit for incorrectly classified instances. Cohn (2003) introduces the concept of (non-uniform) misclassification costs. He makes use of the communicative/semantic distance and proposes a metric for evaluating word sense disambiguation performance using the Receiver Operating Characteristics curve that takes the misclassification costs into account. Bruce and Wiebe (1998) analyze the agreement among human judges for the purpose of formulating a refined and more reliable set of sense tags. Their method is based on statistical analysis of interannotator confusion matrices. An extended study is given in (Bruce and Wiebe, 1999).

4 Conclusion

The usefulness of a semantic resource depends on two aspects:

- reliability of the annotation
- information gain from the annotation.

In practice, each semantic resource emphasizes one aspect: OntoNotes, e.g., guarantees reliability, whereas the WordNet-annotated corpora seek to convey as much semantic nuance as possible. To the best of our knowledge, there has been no exact measure for the optimization, and the usefulness of a given resource can only be assessed when it is finished and used in applications. We propose the *reliable information gain*, a measure based on information theory and on the analysis of interannotator confusion matrices for each word entry, that can be continually applied *during* the creation of a semantic resource, and that provides automatic feedback about the granularity of the used tagset. Moreover, the computed information about the amount of *expected tagging confusion* is also used in evaluation of automatic classifiers.

Acknowledgments

This work has been supported by the Czech Science Foundation projects GK103/12/G084 and P406/2010/0875 and partly by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Ministry of Education, Youth and Sports of the Czech Republic).

We thank our friends from Masaryk University in Brno for providing the annotation infrastructure and for their permanent technical support. We thank Patrick Hanks for his CPA method, for the original PDEV development, and for numerous discussions about the semantics of English verbs. We also thank three anonymous reviewers for their valuable comments.

References

Roni Ben Aharon, Idan Szpektor, and Ido Dagan. 2010. Generating entailment rules from FrameNet. In *Proceedings of the ACL 2010 Conference Short Papers.*, pages 241–246, Uppsala, Sweden.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank II style. Technical report, University of Pennsylvania.

Susan Windisch Brown, Travis Rood, and Martha Palmer. 2010. Number or nuance: Which factors restrict reliable word sense annotation? In *LREC*, pages 3237–3243. European Language Resources Association (ELRA).

Rebecca F. Bruce and Janyce M. Wiebe. 1998. Word-sense distinguishability and inter-coder agreement. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP '98)*, pages 53–60. Granada, Spain, June.

Rebecca F. Bruce and Janyce M. Wiebe. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2):187–205.

Silvie Cinková, Martin Holub, Adam Rambousek, and Lenka Smejkalová. 2012. A database of semantic clusters of verb usages. In *Proceedings of the LREC '2012 International Conference on Language Resources and Evaluation*. To appear.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Trevor Cohn. 2003. Performance metrics for word sense disambiguation. In *Proceedings of the Australasian Language Technology Workshop 2003*, pages 86–93, Melbourne, Australia, December.

Mona T. Diab. 2004. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 303–310. Barcelona, Spain. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore, August. Association for Computational Linguistics.

Katrin Erk. 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*, pages 17–26, Uppsala, Sweden, July. Association for Computational Linguistics.

Christiane Fellbaum, Joachim Grabowski, and Shari Landes. 1997. Analysis of a hand-tagging task. In *Proceedings of the ACL/Siglex Workshop*, Somerset, NJ.

Christiane Fellbaum, J. Grabowski, and S. Landes. 1998. Performance and confidence in a semantic annotation task. In *WordNet: An Electronic Lexical Database*, pages 217–238. Cambridge (Mass.): The MIT Press., Cambridge (Mass.).

Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs, and Susanne Wolf. 2001. Manual and automatic semantic annotation with WordNet.

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

Charles J. Fillmore and B. T. S. Atkins. 1994. Starting where the dictionaries stop: The challenge for computational lexicography. In *Computational Approaches to the Lexicon*, pages 349–393. Oxford University Press.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Francaise de linguistique applique*, 10(2).

Patrick Hanks. forthcoming. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Aleš Horák, Adam Rambousek, and Piek Vossen. 2008. A distributed database system for developing ontological and lexical resources in harmony. In *9th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Berlin: Springer.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passoneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30. European Language Resources Association (ELRA).

Julia Jorgensen. 1990. The psycholinguistic reality of word senses. *Journal of Psycholinguistic Research*, (19):167–190.

Adam Kilgarriff. 1997. "I don't believe in word senses". *Computers and the Humanities*, 31(2):91–113.

Ramesh Krishnamurthy and Diane Nicholls. 2000. Peeling an onion: The lexicographer's experience of manual sense tagging. *Computers and the Humanities*, 34:85–97.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning CoNLLX 06*, pages 216–220. Association for Computational Linguistics.

Adam Meyers, Ruth Reeves, and Catherine Macleod. 2008. NomBank v 1.0.

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. 1993a. A semantic concordance. In *Proceedings of ARPA Workshop on Human Language Technology*.

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. 1993b. A semantic concordance. In *Proceedings of ARPA Workshop on Human Language Technology*.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 105–112, Sydney, Australia.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1).

Rebecca J. Passonneau, Ansaf Salleb-Aoussi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of PolysemousWords by multiple annotators. In *LREC Proceedings*, pages 3244–3249, Valetta, Malta.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Anna Rumshisky, M. Verhagen, and J. Moszkowicz. 2009. The holy grail of sense definition: Creating a Sense-Disambiguated corpus from scratch. Pisa, Italy.

Michael Rundell. 2002. *Macmillan English Dictionary for advanced learners*. Macmillan Education.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. ICSI, University of Berkeley, September.

Beatrice Santorini. 1990. Part-of-Speech tagging guidelines for the penn treebank project. *University of Pennsylvania 3rd Revision 2nd Printing*, (MS-CIS-90-47):33.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

John Sinclair, Patrick Hanks, and et al. 1987. *Collins Cobuild English Dictionary for Advanced Learners 4th edition published in 2003*. HarperCollins Publishers 1987, 1995, 2001, 2003 and Collins A–Z Thesaurus 1st edition first published in 1995. HarperCollins Publishers 1995.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Describing English Language. Oxford University Press.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2011. OntoNotes release 4.0.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. 2009. A machine learning approach to textual entailment recognition. *Natural Language Engineering*, 15(4):551–582.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(November 2009):105–151.