Developing a large semantically annotated corpus

Valerio Basile, Johan Bos, Kilian Evang, Noortje Venhuizen

Developing a large semantically annotated corpus

Valerio Basile, Johan Bos, Kilian Evang, Noortje Venhuizen

Center for Language and Cognition Groningen (CLCG)
University of Groningen
The Netherlands
{v.basile, johan.bos, k.evang, n.j.venhuizen}@rug.nl

Abstract

What would be a good method to provide a large collection of semantically annotated texts with formal, deep semantics rather than shallow? We argue that a bootstrapping approach comprising state-of-the-art NLP tools for parsing and semantic interpretation, in combination with a wiki-like interface for collaborative annotation of experts, and a *game with a purpose* for crowdsourcing, are the starting ingredients for fulfilling this enterprise. The result is a semantic resource that anyone can edit and that *integrates* various phenomena, including predicate-argument structure, scope, tense, thematic roles, rhetorical relations and presuppositions, into a single semantic formalism: Discourse Representation Theory. Taking *texts* rather than sentences as the units of annotation results in deep semantic representations that incorporate discourse structure and dependencies. To manage the various (possibly conflicting) annotations provided by experts and non-experts, we introduce a method that stores "Bits of Wisdom" in a database as stand-off annotations.

Keywords: annotation, semantics, discourse

1. Introduction

Various semantically annotated corpora of reasonable size exist nowadays, including PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998), and the Penn Discourse TreeBank (Prasad et al., 2005). However, efforts that combine various levels of annotation into one formalism are rare. One example is OntoNotes (Hovy et al., 2006), a resource comprising syntax (Penn Treebank style), predicate-argument structure (based on PropBank), word senses, and co-reference. Yet, all of the aforementioned resources lack a level of formally grounded "deep" semantic representation that combines various layers of semantic annotation.

We describe an ongoing effort to fill this gap: the Groningen Meaning Bank (GMB) project. The aim of this project is to provide a large collection of semantically annotated English texts with formal rather than shallow semantics. One of its objectives is to *integrate* phenomena into a single formalism, instead of covering single phenomena in an isolated way. This will provide a better handle on explaining dependencies between various ambiguous linguistic phenomena. Another objective is to annotate *texts*, not isolated sentences (as in ordinary treebanks), which allows us to deal with, for example, ambiguities on the sentence level that require the discourse context for resolving them.

Manually annotating a comprehensive corpus with gold-standard semantic representations is obviously a hard and time-consuming task. Therefore, we use a sophisticated bootstrapping approach. We employ existing NLP tools to get a reasonable approximation of the target annotations to start with. Then we gather and apply *Bits of Wisdom*: pieces of information coming from both experts (linguists) and crowd sourcing methods that help us in deciding how to resolve ambiguities. This will allow us to improve our data-driven NLP machinery and produce improved annotations — and so on.

This paper is organised as follows. First we outline our annotation method, which we dub *human-aided machine annotation*. We illustrate the pipeline of NLP components that we employ, and give some background on the formal semantic representations that we produce. Then we show how we apply *Bits of Wisdom*, BOWs for short, to the annotation and present the first results.

2. Human-Aided Machine Annotation

Our corpus annotation method combines conventional approaches with modern techniques. We use stand-off annotations (based on off-set character positions in the raw text) and automatically produce the annotations from the raw texts to be annotated. This is done by a fairly traditional pipeline of NLP components. The output of the final component is a meaning representation based on Discourse Representation Theory with rhetorical relations, our target annotation for texts. At certain points in the pipeline, the intermediate result may be adjusted by *Bits of Wisdom*, the working of which is explained in detail in section 3.

2.1. Levels of Annotation

The pipeline used for constructing meaning representations is a cascade of various components, most of them provided by the C&C tools and Boxer (Curran et al., 2007). This software, trained and developed on the Penn Treebank, shows high coverage for texts in the newswire domain (up to 98%), is robust and fast, and therefore suitable for producing approximations to gold-standard annotations. The pipeline consists of the following steps:

1. token/sentence boundary detection
2. part-of-speech tagging
3. named entity tagging
4. supertagging (with CCG categories)
5. parsing (syntactic analysis)
6. boxing (semantic analysis)

Table 1: Integrating linguistic information in the GMB

Level	Source	DRS encoding example
POS tag	Penn (Miltsakaki et al., 2004)	
named entity	ENE (Sekine et al., 2002)	`named(X,'John','Person')`
word senses	WordNet (Fellbaum, 1998)	`pred(X,loon,n,2)`
thematic roles	VerbNet (Kipper et al., 2008)	`rel(E,X,'Agent')`
syntax	CCG (Steedman, 2001)	
semantics	DRT (Kamp and Reyle, 1993)	`drs(Referents,Conditions)`
rhetorical relations	SDRT (Asher, 1993)	`rel(DRS1,DRS2,because)`

The semantic analysis is currently carried out by Boxer and will be extended with various external resolution components. These include: scope resolution, anaphora resolution, presupposition projection, assigning thematic roles, word sense disambiguation, discourse segmentation and determining rhetorical relations.

Table 1 shows how the various levels of annotation are integrated in the Groningen Meaning Bank. The linguistic levels of the GMB are, in order of analysis depth: part of speech tags (Penn tagset); named entities (roughly based on (Sekine et al., 2002)); word senses (WordNet); thematic roles (VerbNet); syntactic structure (Combinatory Categorial Grammar, CCG); semantic representations, including events and tense, and rhetorical relations (DRT). Even though we talk about different levels here, they are all connected to each other and integrated in a single formalism: Discourse Representation Theory.

2.2. Discourse Representation Structures

Discourse Representation Theory (DRT) (Kamp and Reyle, 1993) is a widely accepted theory of meaning representation. As the goal of the Groningen Meaning Bank is to provide deep semantic annotations, DRT is in particular suitable because it is designed to incorporate various linguistic phenomena, including the interpretation of pronouns, temporal expressions and plural entities. DRT is based around Discourse Representation Structures (DRSs), which are recursive formal meaning structures that have a model-theoretic interpretation. This interpretation can be given directly (Kamp and Reyle, 1993) or via a translation into first-order logic (Muskens, 1996). This property is not only interesting from a theoretical point of view, but also from a practical perspective, because it permits the use of efficient existing inference engines (e.g. theorem provers and model builders) developed by the automated deduction community.

The aim of the Groningen Meaning Bank is to provide fully resolved semantic representations. This inspires the adoption of well-known extensions to the standard theory of DRT to include neo-Davidsonian events (with VerbNet roles (Kipper et al., 2008)), presuppositions (van der Sandt, 1992) and rhetorical relations (Asher, 1993). The latter feature is part and parcel of SDRT (Segmented Discourse Representation Theory), a theory that enriches DRT's semantics with a precise dynamic semantics for rhetorical relations (Asher and Lascarides, 2003). In the GMB, rhetorical relations are represented as relations between DRSs, which may in turn be embedded in another DRS. We distinguish different types of rhetorical relations (both coordinating and subordinating relations), and separate the presuppositions of the discourse from the asserted content. Future work may include the representation of ambiguities by adding some underspecification mechanisms to the formalism.

The trademark of the Groningen Meaning Bank is that it provides the information of various layers of meaning within a single representation format: a DRS. Figure 1 shows the semantic representation of an example text from the GMB.

3. Bits of Wisdom

In order to improve and refine the analysis provided by the tools, their output (which takes the form of text and XML files in various formats) may be adjusted by human expert annotators, crowd sourcing activities or external software components. To prevent these various sources of annotation from stepping on each other's toes, we defined the basic unit of annotation input as what we call a *Bit of Wisdom* (BOW). A BOW is represented as a database entry that gives advice on a particular linguistic interpretation decision. BOWs can inform where a sentence boundary occurs, where a token starts and ends, what part of speech is assigned to a token, what thematic role is played in an event, what sense of a word is meant, and so on.

3.1. Types of BOWs

We distinguish different types of BOWs, each type with a fixed set of integer or string arguments that contain the actual information. Currently, we work with two types of BOWs:

(i) $boundary(n, level, polarity)$ where n is a character offset into the text, $level \in$ {token,sentence} and $polarity \in \{+, -\}$, meaning that there is a/there is no token/sentence boundary at n,

(ii) $tag(l, r, type, tag)$ where l, r are character offsets, $type$ is a tag type such as POS, word sense, NE or supertag, and tag is a tag of that tag type, meaning that the token between these offsets should carry this tag.

Bits of Wisdom may be applied at different points in the pipeline, e.g., after sentence and token boundary detection in the case of type (i), or after the relevant tagging step in the case of type (ii).

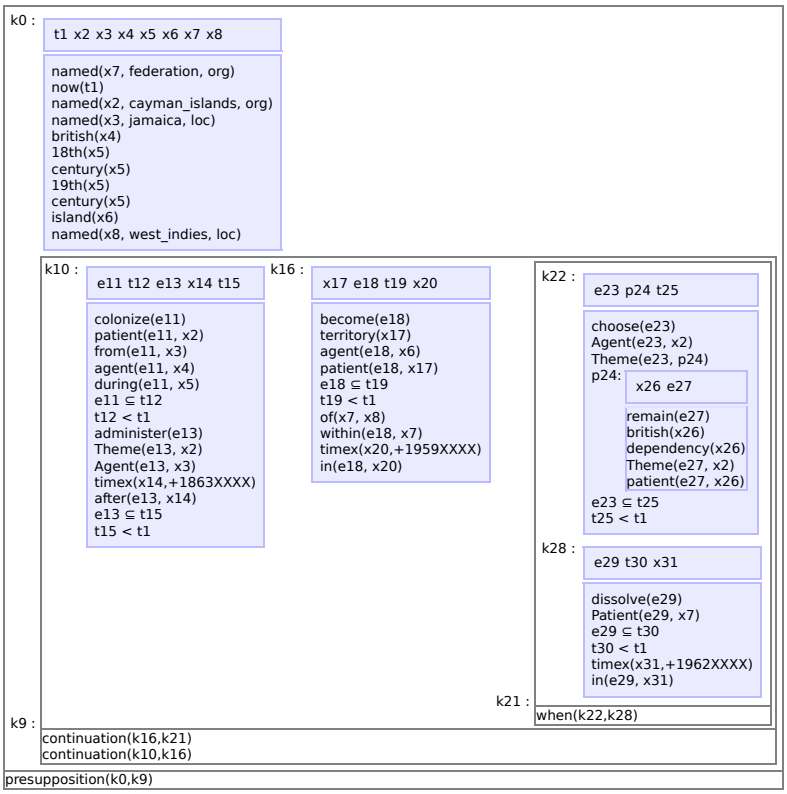

Figure 1: DRS for the text "The Cayman Islands were colonized from Jamaica by the British during the 18th and 19th centuries and were administered by Jamaica after 1863. In 1959, the islands became a territory within the Federation of the West Indies. When the Federation dissolved in 1962, the Cayman Islands chose to remain a British dependency." (document 55/0688 in the GMB)

Applying a BOW amounts to making the minimal set of changes to the annotation required to make it consistent with the BOW while keeping the overall annotation well-formed. This involves, for example, securing a complete tokenization of the text, avoiding multiple tags of the same type for the same token and creating a token boundary whenever a sentence boundary occurs, or removing a sentence boundary in case a token boundary is removed. In the unusual event that a BOW cannot be applied because, for example, it refers to a token that does not exist anymore due to changed boundaries, the BOW is ignored.

We expect that most of the automatically generated syntactic analyses can be corrected by applying BOWs with the correct lexical category, because CCG is a lexicalised theory of grammar.

3.2. Sources of BOWs

There are currently two sources of BOWs, namely a group of experts editing the annotation in a wiki-like fashion through a Web-based tool called GMB Explorer, and a group of non-experts that provide information for the lower levels of annotation decisions (e.g. word senses) by way of a 'Game with a Purpose', called *Wordrobe*. In this way, we gather bits of both expert wisdom and collective wisdom. A third source of BOWs in future work will be the use of external components, like state-of-the-art WSD, co-reference resolution or semantic role labelling systems.

The first source of BOWs, the GMB Explorer, allows users to make *changes* to the annotation, such as splitting a token in two, merging an erroneously split sentence back together, or correcting a tag on a token (Basile et al., forthcoming 2012). All changes are stored as BOWs, so they are independent of the previous state of the annotation and can be applied selectively and in any order. We currently apply all expert BOWs for a particular document each time the pipeline is run on that document, in chronological order with the most recent BOW last. This way, experts can freely edit the latest state of the annotation, and the complexity of the BOW application process is linear in the number of BOWs. GMB Explorer makes the collaborative annotation process transparent through global and per-document newsfeeds of BOWs, similar to the "recent changes" and "history" feature of wikis. This is exemplified in Figure 2.

The second source for BOWs, the game with a purpose, is similar to successful initiatives like *Phrase Detectives* (Chamberlain et al., 2008) and *Jeux de Mots* (Artignan et al., 2009). It collects answers from non-expert players to problems such as choosing the sense of a word in a given context from a list of definitions. Once the same answer has been given by a significant majority of the players, it is likely to be correct and a BOW is automatically generated.

3.3. Judging BOWs

It is important to stress that a single BOW isn't necessarily correct, but rather gives an opinion or prediction of a certain linguistic interpretation – albeit from an authoritative source, thus with a relatively high reliability. As a consequence of the defeasibility of BOWs, a particular annotation

time	user	document	cat0	cat1	cat2	content
2012-02-15 14:01:51	johan.bos	93/0218	BOWI	tag	ne	token *London* at <66,72> has ne tag: Organization
2012-02-15 14:01:51	johan.bos	93/0218	BOWI	tag	ne	token *Spears* at <42,48> has ne tag: Person
2012-02-15 14:01:51	johan.bos	93/0218	BOWI	tag	ne	token *Thing* at <468,473> has ne tag: Artifact
2012-02-15 13:52:02	johan.bos	94/0231	BOWI	tag	ne	token *almost* at <205,211> has ne tag: O
2012-02-15 11:25:49	noortje	66/0653	BOWI	tag	pos	token *suspect* at <48,55> has pos tag: NN
2012-02-15 11:10:06	noortje	01/0630	BOWI	tok	sentence	sentence boundary at 184: *4.9.* | *Officials say*
2012-02-15 11:10:05	noortje	01/0630	BOWI	tok	token	token boundary at 182: *4.9* .
2012-02-15 03:23:07	johan.bos	53/0517	BOWI	tok	token	no token boundary at 194: *Oct*_.
2012-02-15 03:23:07	johan.bos	53/0517	BOWI	tok	sentence	no sentence boundary at 196: *Oct.* _ *1,*

Figure 2: Global newsfeed of recently added BOWs in GMB Explorer.

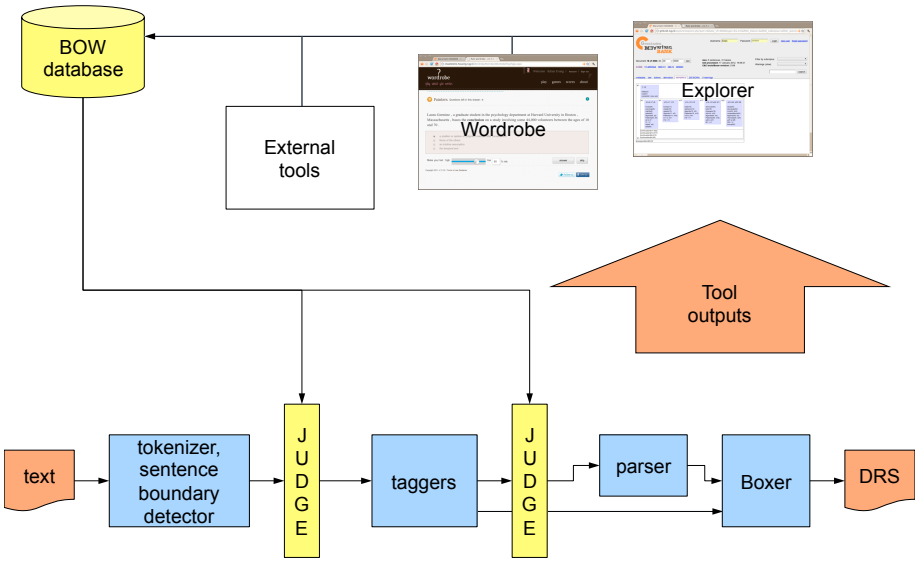

Figure 3: Graphical representation of the workflow for constructing the GMB.

choice can be supported by some Bits of Wisdom, and rejected by others, which may arise from various sources with different reliabilities. Whenever there is disagreement between BOWs, a *judge component* decides how to interpret them. The decisions of the judge may be consistent with the respective NLP component's output, or conflict with it. In the latter case the output is adjusted. This process takes place directly after each component in the pipeline, so subsequent components take advantage of previous decisions, producing higher quality output. The judge component that resolves possible conflicts between different types of BOWs is currently as simple as always applying expert BOWs in the end, giving them priority. We will be investigating other judging methods as we add external software components as a third source of BOWs.

4. Constructing the GMB

4.1. Workflow

In the previous sections we have introduced various techniques used for creating the Groningen Meaning Bank. Combining these techniques results in the worflow depicted in Figure 3. It consists of the NLP pipeline, interleaved with judge components, and a feedback loop. The feedback loop works as follows: the intermediate (e.g. tokens, tags, parse tree) and final outputs (DRS) of the pipeline are visualized in Explorer and used for generating questions for Wordrobe. The BOWs provided by these two sources, as well as annotations created by external tools, are stored in the database and are then judged and applied in subsequent runs of the pipeline. Currently, BOWs can be applied at two points in de pipeline, namely after the token/sentence boundary detector and after the tagging of POS, named entities and CCG categories. The process is orchestrated by the tool *GNU make* and a daemon process that schedules the reprocessing of individual documents as new BOWs are added to the database.

4.2. Data

We believe that a semantically annotated corpus would be extremely valuable for data-driven semantic analysis and future developments in computational semantics, in the same way that treebanks have played a crucial role for the development of robust parsers. Therefore, a corpus developed primarily for research purposes ought to be widely and easily available to researchers in the field. As a consequence, the GMB only comprises texts whose distribution isn't subject to copyright restrictions. Included are newswire texts from Voice of America, country descrip-

tions from the CIA Factbook, a collection of texts from the open ANC (Ide et al., 2010) and Aesop's fables. All of these documents are in the public domain.

Size and quality are factors that influence the usefulness of annotated resources. Since one of the things we have in mind is the use of statistical techniques in natural-language generation, the corpus should be sufficiently large. We aim to provide a trade-off between quality and quantity, with a process that increases the size of the corpus and improves the annotation accuracy in each stable release of the GMB.

4.3. Availability

A first release comprising 1,000 texts with 4,239 sentences and 82,752 tokens is available at `http://gmb.let.rug.nl`. A further 70K texts with more than 1M sentences and 31M tokens has so far been collected, ready for inclusion in future releases. The current development version contains 5,000 texts and is accessible online through GMB Explorer, where registered users can contribute BOWs. We plan to release a first version of the annotation game *Wordrobe* within the coming months.

5. Conclusions

In this paper we introduced human-aided machine annotation, a method for developing a large semantically annotated corpus, as applied in the Groningen Meaning Bank. The method uses state-of-the art NLP tools in combination with human input in the form of *Bits of Wisdom*.

As the goal of the Groningen Meaning Bank is to create a gold standard for meaning representations, future work will include quantifying the degree to which the gold standard is reached for a certain representation in terms of the Bits of Wisdom applied to the representation. Our working hypothesis is that the more BOWs are applied, the closer the representation comes to a gold standard.

6. References

Guillaume Artignan, Mountaz Hascoët, and Mathieu Lafourcade. 2009. Multiscale visual analysis of lexical networks. *Information Visualisation, International Conference on*, 0:685–690.

N. Asher and A. Lascarides. 2003. *Logics of conversation*. Studies in natural language processing. Cambridge University Press.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Proceedings of the Conference*, pages 86–90, Université de Montréal, Montreal, Quebec, Canada.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. forthcoming 2012. A platform for collaborative semantic annotation. In *Proceedings of EACL 2012*.

John Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 375–380. College Publications.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

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*, pages 57–60, Stroudsburg, PA, USA.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: a community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Stroudsburg, PA, USA.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *In Proceedings of LREC 2004*, pages 2237–2240.

Reinhard Muskens. 1996. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy*, 19:143–186.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proc. of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

S. Sekine, K. Sudo, and C. Nobata. 2002. Extended named entity hierarchy. In *Proceedings of LREC*, volume 2.

Satoshi Sekine. 2010. Sekine's extended named entity hierarchy. `http://nlp.cs.nyu.edu/ene/`.

Mark Steedman. 2001. *The Syntactic Process*. The MIT Press.

Rob A. van der Sandt. 1992. Presupposition Projection as Anaphora Resolution. *Journal of Semantics*, 9:333–377.