Coarse Lexical Semantic Annotation with Supersenses: An Arabic Case Study

Nathan Schneider†  Behrang Mohit*  Kemal Oflazer*  Noah A. Smith†
School of Computer Science, Carnegie Mellon University
*Doha, Qatar  †Pittsburgh, PA 15213, USA
{nensch@cs., behrang@, ko@cs., nasmith@cs.}cmu.edu

Abstract

“Lightweight” semantic annotation of text calls for a simple representation, ideally without requiring a semantic lexicon to achieve good coverage in the language and domain. In this paper, we repurpose WordNet’s supersense tags for annotation, developing specific guidelines for nominal expressions and applying them to Arabic Wikipedia articles in four topical domains. The resulting corpus has high coverage and was completed quickly with reasonable inter-annotator agreement.

1 Introduction

The goal of “lightweight” semantic annotation of text, particularly in scenarios with limited resources and expertise, presents several requirements for a representation: simplicity; adaptability to new languages, topics, and genres; and coverage. This paper describes coarse lexical semantic annotation of Arabic Wikipedia articles subject to these constraints. Traditional lexical semantic representations are either narrow in scope, like named entities, or make reference to a full-fledged lexicon/ontology, which may insufficiently cover the language/domain of interest or require prohibitive expertise and effort to apply. We therefore turn to supersense tags (SSTs), 40 coarse lexical semantic classes (25 for nouns, 15 for verbs) originating in WordNet. Previously these served as groupings of English lexicon

1Some ontologies like those in Sekine et al. (2002) and BBN Identifinder (Bikel et al., 1999) include a large selection of classes, which tend to be especially relevant to proper names.

2E.g., a WordNet (Fellbaum, 1998) sense annotation effort reported by Passonneau et al. (2010) found considerable inter-annotator variability for some lexemes; FrameNet (Baker et al., 1998) is limited in coverage, even for English; and PropBank (Kingsbury and Palmer, 2002) does not capture semantic relationships across lexemes. We note that the Omega ontology (Philpot et al., 2003) has been used for fine-grained cross-lingual annotation (Hovy et al., 2006; Dorr et al., 2010).

3Note that work in supersense tagging used text with fine-grained sense annotations that were then coarsened to SSTs.

4The noun/verb distinction might prove problematic in some languages.
Table 1: Snapshot of the supersense-annotated data. The 7 article titles (translated) in each domain, with total counts of sentences, tokens, and supersense mentions. Overall, there are 2,219 sentences with 65,452 tokens and 23,239 mentions (1.3 tokens/mention on average). Counts exclude sentences marked as problematic and mentions marked ?.

3 Arabic Wikipedia Annotation

The annotation in this work was on top of a small corpus of Arabic Wikipedia articles that had already been annotated for named entities (Mohit et al., 2012). Here we use two different annotators, both native speakers of Arabic attending a university with English as the language of instruction.

Data & procedure. The dataset (table 1) consists of the main text of 28 articles selected from the topical domains of history, sports, science, and technology. The annotation task was to identify and categorize mentions, i.e., occurrences of terms belonging to noun supersenses. Working in a custom, browser-based interface, annotators were to tag each relevant token with a supersense category by selecting the token and typing a tag symbol. Any token could be marked as continuing a multiword unit by typing <. If the annotator was ambivalent about a token they were to mark it with the ? symbol. Sentences were pre-tagged with suggestions where possible.

Annotations noted obvious errors in sentence splitting and grammar so ill-formed sentences could be excluded.

Training. Over several months, annotators alternately annotated sentences from 2 designated articles of each domain, and reviewed the annotations for consistency. All tagging conventions were developed collaboratively by the author(s) and annotators during this period, informed by points of confusion and disagreement. WordNet and SemCor were consulted as part of developing the guidelines, but not during annotation itself so as to avoid complicating the annotation process or overfitting to WordNet’s idiosyncrasies. The training phase ended once inter-annotator mention \( F_1 \) had reached 75%.

Suggestions came from the previous named entity annotation of PERSONS, organizations (GROUP), and LOCATIONS, as well as heuristic lookup in lexical resources—Arabic WordNet entries (Elkateb et al., 2006) mapped to English WordNet, and named entities in OntoNotes (Hovy et al., 2006).
The complete supersense tagset for nouns; each tag is briefly described by its symbol, NAME, short description, and examples. Some examples and longer descriptions have been omitted due to space constraints.

### Domain-Specific Elaborations

- **Science**: chemicals, molecules, atoms, and subatomic particles are tagged as SUBSTANCE
- **Sports**: championships/tournaments are EVENTS
- **(Information) Technology**: Software names, kinds, and components are tagged as COMMUNICATION (e.g. kernel, version, distribution, environment). A connection is a RELATION; project, support, and a configuration are tagged as COGNITION; development and collaboration are ACTS.
- **Arabic conventions**: Masdar constructions (verbal nouns) are treated as nouns. Anaphora are not tagged.

Figure 2: Above: The complete supersense tagset for nouns; each tag is briefly described by its symbol, NAME, short description, and examples. Some examples and longer descriptions have been omitted due to space constraints. Below: A few domain- and language-specific elaborations of the general guidelines.
Main annotation. After training, the two annotators proceeded on a per-document basis: first they worked together to annotate several sentences from the beginning of the article, then each was independently assigned about half of the remaining sentences (typically with 5–10 shared to measure agreement). Throughout the process, annotators were encouraged to discuss points of confusion with each other, but each sentence was annotated in its entirety and never revisited. Annotation of 28 articles required approximately 100 annotator-hours. Articles used in pilot rounds were re-annotated from scratch.

Analysis. Figure 3 shows the distribution of SSTs in the corpus. Some of the most concrete tags—BODY, ANIMAL, PLANT, NATURAL OBJECT, and FOOD—were barely present, but would likely be frequent in life sciences domains. Others, such as MOTIVE, POSSESSION, and SHAPE, are limited in scope.

To measure inter-annotator agreement, 87 sentences (2,774 tokens) distributed across 19 of the articles (not including those used in pilot rounds) were annotated independently by each annotator. Inter-annotator mention $F_1$ (counting agreement over entire mentions and their labels) was 70%. Excluding the 1,397 tokens left blank by both annotators, the token-level agreement rate was 71%, with Cohen’s $\kappa = 0.69$, and token-level $F_1$ was 83%.\(^7\)

We also measured agreement on a tag-by-tag basis. For 8 of the 10 most frequent SSTs (figure 3), inter-annotator mention $F_1$ ranged from 73% to 80%. The two exceptions were QUANTITY at 63%, and COGNITION (probably the most heterogeneous category) at 49%. An examination of the confusion matrix reveals four pairs of supersense categories that tended to provoke the most disagreement: COMMUNICATION/COGNITION, ACT/COGNITION, ACT/PROCESS, and ARTIFACT/COMMUNICATION.

Figure 3: Distribution of supersense mentions by domain (left), and counts for tags occurring over 800 times (below). (Counts are of the union of the annotators’ choices, even when they disagree.)

| tag         | num | tag         | num |
|-------------|-----|-------------|-----|
| ACT (†)     | 3473| LOCATION (G) | 1583|
| COMMUNICATION (C) | 3007| GROUP (L) | 1501|
| PERSON (F)  | 2650| TIME (T) | 1407|
| ARTIFACT (A) | 2164| SUBSTANCE (S) | 1291|
| COGNITION ( perpetrated | 1672| QUANTITY (Q) | 1022|

The last is exhibited for the first mention in figure 1, where one annotator chose ARTIFACT (referring to the physical book) while the other chose COMMUNICATION (the content). Also in that sentence, annotators disagreed on the second use of university (ARTIFACT vs. GROUP). As with any sense annotation effort, some disagreements due to legitimate ambiguity and different interpretations of the tags—especially the broadest ones—are unavoidable.

A “soft” agreement measure (counting as matches any two mentions with the same label and at least one token in common) gives an $F_1$ of 79%, showing that boundary decisions account for a major portion of the disagreement. E.g., the city Fez, Morocco (figure 1) was tagged as a single LOCATION by one annotator and as two by the other. Further examples include the technical term ‘thin client’, for which one annotator omitted the adjective; and ‘World Cup Football Championship’, where one annotator tagged the entire phrase as an EVENT while the other tagged ‘football’ as a separate ACT.

4 Conclusion

We have codified supersense tags as a simple annotation scheme for coarse lexical semantics, and have shown that supersense annotation of Arabic Wikipedia can be rapid, reliable, and robust (about half the tokens in our data are covered by a nominal supersense). Our tagging guidelines and corpus are available for download at http://www.ark.cs.cmu.edu/ArabicSST/.

Acknowledgments

We thank Nourhen Feki and Sarah Mustafa for assistance with annotation, as well as Emad Mohamed, CMU ARK members, and anonymous reviewers for their comments. This publication was made possible by grant NPRP-08-485-1-083 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.
References

Giuseppe Attardi, Stefano Dei Rossi, Giulia Di Pietro, Alessandro Lenci, Simonetta Montemagni, and Maria Simi. 2010. A resource and tool for super-sense tagging of Italian texts. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta, May. European Language Resources Association (ELRA).

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL ’98), pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

D. M. Bikel, R. Schwartz, and R. M. Weischedel. 1999. An algorithm that learns what’s in a name. Machine Learning, 34(1).

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 594–602, Sydney, Australia, July. Association for Computational Linguistics.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pages 168–175, Sapporo, Japan, July.

James R. Curran. 2005. Supersense tagging of unknown nouns using semantic similarity. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL’05), pages 26–33, Ann Arbor, Michigan, June.

Bonnie J. Dorr, Rebecca J. Passonneau, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith J. Miller, Teruko Mitamura, Owen Rambow, and Advaith Siddharthan. 2010. Interlingual annotation of parallel text corpora: a new framework for annotation and evaluation. Natural Language Engineering, 16(03):197–243.

Sabri Elkateb, William Black, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a WordNet for Arabic. In Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC 2006), pages 29–34, Genoa, Italy.

Christiane Fellbaum. 1990. English verbs as a semantic net. International Journal of Lexicography, 3(4):278–301, December.

Christiane Fellbaum, editor. 1998. WordNet: an electronic lexical database. MIT Press, Cambridge, MA.

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 (HLT-NAACL), pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02), Las Palmas, Canary Islands, May.

Xin Li and Dan Roth. 2002. Learning question classifiers. In Proceedings of the 19th International Conference on Computational Linguistics (COLING’02), pages 1–7, Taipei, Taiwan, August. Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In Proceedings of the Workshop on Human Language Technology (HLT ’93), pages 303–308, Plainsboro, NJ, USA, March. Association for Computational Linguistics.

George A. Miller. 1990. Nouns in WordNet: a lexical inheritance system. International Journal of Lexicography, 3(4):245–264, December.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), pages 162–173, Avignon, France, April. Association for Computational Linguistics.

Gerhard Paass and Frank Reichartz. 2009. Exploiting semantic constraints for estimating supersenses with CRFs. In Proceedings of the Ninth SIAM International Conference on Data Mining, pages 485–496, Sparks, Nevada, USA, May. Society for Industrial and Applied Mathematics.

Rebecca J. Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta, May. European Language Resources Association (ELRA).
Andrew G. Philpot, Michael Fleischman, and Eduard H. Hovy. 2003. Semi-automatic construction of a general purpose ontology. In Proceedings of the International Lisp Conference, New York, NY, USA, October.

Davide Picca, Alfio Massimiliano Gliozzo, and Massimiliano Ciaramita. 2008. Supersense Tagger for Italian. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), pages 2386–2390, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Davide Picca, Alfio Massimiliano Gliozzo, and Simone Campora. 2009. Bridging languages by SuperSense entity tagging. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pages 136–142, Suntec, Singapore, August. Association for Computational Linguistics.

Likun Qiu, Yunfang Wu, Yanqiu Shao, and Alexander Gelbukh. 2011. Combining contextual and structural information for supersense tagging of Chinese unknown words. In Computational Linguistics and Intelligent Text Processing: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing’11), volume 6608 of Lecture Notes in Computer Science, pages 15–28. Springer, Berlin.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02), Las Palmas, Canary Islands, May.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O’Connor, and Tom Wasow. 2004. Animacy encoding in English: why and how. In Bonnie Webber and Donna K. Byron, editors, ACL 2004 Workshop on Discourse Annotation, pages 118–125, Barcelona, Spain, July. Association for Computational Linguistics.