Coarse Lexical Semantic Annotation with Supersenses: An Arabic Case Study

Nathan Schneider
Carnegie Mellon University

Behrang Mohit
Carnegie Mellon University, behrang@cmu.edu

Kemal Oflazer
Carnegie Mellon University, ko@andrew.cmu.edu

Noah A. Smith
Carnegie Mellon University, nasmith@cs.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/lti

Part of the Computer Sciences Commons

Published In
Proceedings of the Annual Meeting of the Association for Computational Linguistics, 253-258.
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

{nschneid@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 entries, but here we have repurposed them as target labels for direct human annotation.

Part of the earliest versions of WordNet, the supersense categories (originally, “lexicographer classes”) were intended to partition all English noun and verb senses into broad groupings, or semantic fields (Miller, 1990; Fellbaum, 1990). More recently, the task of automatic supersense tagging has emerged for English (Ciaramita and Johnson, 2003; Curran, 2005; Ciaramita and Altun, 2006; Paaß and Reichartz, 2009), as well as for Italian (Picca et al., 2008; Picca et al., 2009; Attardi et al., 2010) and Chinese (Qiu et al., 2011), languages with WordNets mapped to English WordNet. In principle, we believe supersenses ought to apply to nouns and verbs in any language, and need not depend on the availability of a semantic lexicon. In this work we focus on the noun SSTs, summarized in figure 2 and applied to an Arabic sentence in figure 1.

The first sentence in the “Islamic Golden Age” article is shown in figure 1, with supersense tagging from one of two annotators. The Arabic text is shown left-to-right.

Figure 1: A sentence from the article “Islamic Golden Age,” with the supersense tagging from one of two annotators. The Arabic is shown left-to-right.

The Guinness Book of World Records considers the University of Al-Karaouine in Fez, Morocco, established in the year 859 AD, the oldest university in the world.’

1 Some 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.

2 E.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).

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

4 The noun/verb distinction might prove problematic in some languages.
disambiguating PERSON vs. POSSESSION for the noun principal—and generalize across lexemes on the other—e.g., principal, teacher, and student can all be PERSONS. This lumping property might be expected to give too much latitude to annotators; yet we find that in practice, it is possible to elicit reasonable inter-annotator agreement, even for a language other than English. We encapsulate our interpretation of the tags in a set of brief guidelines that aims to be usable by anyone who can read and understand a text in the target language; our annotators had no prior expertise in linguistics or linguistic annotation.

Finally, we note that ad hoc categorization schemes not unlike SSTs have been developed for purposes ranging from question answering (Li and Roth, 2002) to animacy hierarchy representation for corpus linguistics (Zaenen et al., 2004). We believe the interpretation of the SSTs adopted here can serve as a single starting point for diverse resource engineering efforts and applications, especially when fine-grained sense annotation is not feasible.

2 Tagging Conventions

WordNet’s definitions of the supersenses are terse, and we could find little explicit discussion of the specific rationales behind each category. Thus, we have crafted more specific explanations, summarized for nouns in figure 2. English examples are given, but the guidelines are intended to be language-neutral. A more systematic breakdown, formulated as a 43-rule decision list, is included with the corpus. In developing these guidelines we consulted English WordNet (Fellbaum, 1998) and SemCor (Miller et al., 1993) for examples and synset definitions, occasionally making simplifying decisions where we found distinctions that seemed esoteric or internally inconsistent. Special cases (e.g., multiword expressions, anaphora, figurative language) are addressed with additional rules.

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. Annotators 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 idiosyncracies. The training phase ended once inter-annotator mention F1 had reached 75%.

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 ?.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Sentences</th>
<th>Tokens</th>
<th>Mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>390</td>
<td>13,716</td>
<td>5,149</td>
</tr>
<tr>
<td>Sports</td>
<td>375</td>
<td>11,523</td>
<td>4,162</td>
</tr>
<tr>
<td>Science</td>
<td>1,826</td>
<td>53,655</td>
<td>20,123</td>
</tr>
<tr>
<td>Technology</td>
<td>618</td>
<td>16,992</td>
<td>5,754</td>
</tr>
<tr>
<td>Total</td>
<td>2,219</td>
<td>65,452</td>
<td>23,239</td>
</tr>
</tbody>
</table>

5For example, one rule states that all man-made structures (buildings, rooms, bridges, etc.) are to be tagged as ARTIFACTS.
<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NATURAL OBJECT</strong></td>
<td>natural feature or nonliving object in nature. barrier_reef nest neutron star planet sky fishpond metamorphic_rock Mediterranean cave stepping_stone boulder Orion ember universe</td>
</tr>
<tr>
<td><strong>ARTIFACT</strong></td>
<td>man-made structures and objects bridge restaurant bedroom stage cabinet toaster antitoxin aspirin</td>
</tr>
<tr>
<td><strong>LOCATION</strong></td>
<td>any name of a geopolitical entity, as well as other nouns functioning as locations or regions. Cote_d'Ivoire New_York_City downtown stage left India Newark interior airspace</td>
</tr>
<tr>
<td><strong>PERSON</strong></td>
<td>humans or personified beings; names of social groups (ethnic, political, etc.) that can refer to an individual in the singular. Persian_deity glasscutter mother kibbutznik firstborn worshiper Roosevelt Arab consumer appellant guardsmen Muslim American communist</td>
</tr>
<tr>
<td><strong>GROUP</strong></td>
<td>groupings of people or objects, including: organizations/institutions; followers of social movements collection flock army meeting clergy Mennonite_Chamber trump section health profession peasantry People's_Party U.S._State_Department University_of_California population consulting_firm communism Islam (= set of Muslims)</td>
</tr>
<tr>
<td><strong>SUBSTANCE</strong></td>
<td>a material or substance krypton mocha atom hydrochloric_acid aluminum sand cardboard DNA</td>
</tr>
<tr>
<td><strong>POSSESSION</strong></td>
<td>term for an entity involved in ownership or payment birthday present tax_shelter money loan</td>
</tr>
<tr>
<td><strong>TIME</strong></td>
<td>a temporal point, period, amount, or measurement 10_seconds day Eastern_Time leap_year 2nd_million_BCE 2011 (= year) velocity frequency runtime latency/delay middle_age half_life basketball_season words_per_minute curfew industrial_revolution instant/moment August</td>
</tr>
<tr>
<td><strong>RELATION</strong></td>
<td>relations between entities or quantities ratio scale reverse personal_relation exponential_function angular_position unconnectedness transitivity</td>
</tr>
<tr>
<td><strong>QUANTITY</strong></td>
<td>quantities and units of measure, including cardinal numbers and fractional amounts 7_cm 1.8_million 12_percent/12% volume (= spatial extent) volt real_number square_root digit 90_degrees handful ounce half</td>
</tr>
<tr>
<td><strong>FEELING</strong></td>
<td>subjective emotions indifference wonder murderousness grudge desperation astonishing suffering</td>
</tr>
<tr>
<td><strong>MOTIVE</strong></td>
<td>an abstract external force that causes someone to intend to do something. reason incentive</td>
</tr>
<tr>
<td><strong>COMMUNICATION</strong></td>
<td>information encoding and transmission, except in the sense of a physical object grave Accent Book_of_Common_Prayer alphabet Cree_language onomatopoeia reference concert hotel_bill broadcast television_program discussion contract proposal equation denial sarcasm concerto software</td>
</tr>
<tr>
<td><strong>COGNITION</strong></td>
<td>aspects of mind/thought/knowledge/belief/perception; techniques and abilities; fields of academic study; social or philosophical movements referring to the system of beliefs. Platonism hypothesis logic biomedical_science necromancy hierarchical_structure democracy innovativeness vocational_program woodcraft reference visual_image Islam (= Islamic belief system) dream scientific_method consciousness puzzlement skepticism reasoning design intuition inspiration muscle_memory skill aptitude/talent method sense_of_touch awareness</td>
</tr>
<tr>
<td><strong>STATE</strong></td>
<td>stable states of affairs; diseases and their symptoms symptom reprieve potency poverty altitude sickness tumor fever measles bankruptcy infamy opulence hunger opportunity darkness (= lack of light)</td>
</tr>
<tr>
<td><strong>ATTRIBUTE</strong></td>
<td>characteristics of people/objects that can be judged resilience buxomness virtue immateriality admissibility coincidence valence sophistication simplicity temperature (= degree of hotness) darkness (= dark coloring)</td>
</tr>
<tr>
<td><strong>ACT</strong></td>
<td>things people do or cause to happen; learned professions meddling malpractice faith_healing dismount carnival football</td>
</tr>
</tbody>
</table>
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%.\footnote{Token-level measures consider both the supersense label and whether it begins or continues the mention.}

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.

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


