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*,[1] 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.[2] 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.

أن القياسية للأرقام جينيس كتاب يعتبر
considers book Guinness for-records the-standard that
COMMUNICATION

في جامعة أقدم المغرب فاس في القيروان جامعة
university Al-Karaouine in Fez Morocco oldest university in
ARTIFACT LOCATION GROUP

العالم تم حيث تأسيسها في 859 سنة ميلادي .
the-world where was established in year AD
LOCATION ACT TIME

'*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.'

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.

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.[3] 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.[4] In this work we focus on the noun SSTs, summarized in figure 2 and applied to an Arabic sentence in figure 1.

SSTs both refine and relate lexical items: they capture lexical polysemy on the one hand—e.g.,

[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.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 253–258,
Jeju, Republic of Korea, 8-14 July 2012.

Crusades · Damascus · Ibn Tolun Mosque · Imam Hussein Shrine · Islamic Golden Age · Islamic History · Ummayad Mosque	434s	16,185t	5,859m
Atom · Enrico Fermi · Light · Nuclear power · Periodic Table · Physics · Muhammad al-Razi	777s	18,559t	6,477m
2004 Summer Olympics · Christiano Ronaldo · Football · FIFA World Cup · Portugal football team · Raúl Gonzáles · Real Madrid	390s	13,716t	5,149m
Computer · Computer Software · Internet · Linux · Richard Stallman · Solaris · X Window System	618s	16,992t	5,754m

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 ?.

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.[5] 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.[6] 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 F_1 had reached 75%.

[5] For example, one rule states that all man-made structures (buildings, rooms, bridges, etc.) are to be tagged as ARTIFACTs.

[6] 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).

O **NATURAL OBJECT** 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**

A **ARTIFACT** man-made structures and objects **bridge restaurant bedroom stage cabinet toaster antidote aspirin**

L **LOCATION** 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**

P **PERSON** 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 guardsman Muslim American communist**

G **GROUP** groupings of people or objects, including: organizations/institutions; followers of social movements **collection flock army meeting clergy Mennonite_Church trumpet_section health_profession peasantry People's_Party U.S._State_Department University_of_California population consulting_firm communism Islam** (= set of Muslims)

$ **SUBSTANCE** a material or substance **krypton mocha atom hydrochloric_acid aluminum sand cardboard DNA**

H **POSSESSION** term for an entity involved in ownership or payment **birthday_present tax_shelter money loan**

T **TIME** a temporal point, period, amount, or measurement **10_seconds day Eastern_Time leap_year 2nd_millenium_BC 2011** (= year) **velocity frequency runtime latency/delay middle_age half_life basketball_season words_per_minute curfew industrial_revolution instant/moment August**

= **RELATION** relations between entities or quantities **ratio scale reverse personal_relation exponential_function angular_position unconnectedness transitivity**

Q **QUANTITY** 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**

F **FEELING** subjective emotions **indifference wonder murderousness grudge desperation astonishment suffering**

M **MOTIVE** an abstract external force that causes someone to intend to do something **reason incentive**

C **COMMUNICATION** 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**

^ **COGNITION** 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**

S **STATE** 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)

@ **ATTRIBUTE** characteristics of people/objects that can be judged **resilience buxomness virtue immateriality admissibility coincidence valence sophistication simplicity temperature** (= degree of hotness) **darkness** (= dark coloring)

! **ACT** things people do or cause to happen; learned professions **meddling malpractice faith_healing dismount carnival football_game acquisition engineering** (= profession)

E **EVENT** things that happens at a given place and time **bomb_blast ordeal miracle upheaval accident tide**

R **PROCESS** a sustained phenomenon or one marked by gradual changes through a series of states **oscillation distillation overheating aging accretion/growth extinction evaporation**

X **PHENOMENON** a physical force or something that happens/occurs **electricity suction tailwind tornado effect**

+ **SHAPE** two and three dimensional shapes

D **FOOD** things used as food or drink

B **BODY** human body parts, excluding diseases and their symptoms

Y **PLANT** a plant or fungus

N **ANIMAL** non-human, non-plant life

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.

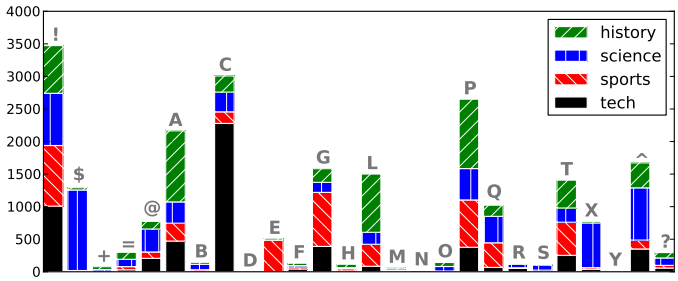

Figure 3: Distribution of supersense mentions by domain (left), and counts for tags occurring over 800 times (below). (Counts are of the *union* of the annotators' choices, even when they disagree.)

tag	*num*	*tag*	*num*
ACT (!)	3473	LOCATION (G)	1583
COMMUNICATION (C)	3007	GROUP (L)	1501
PERSON (P)	2650	TIME (T)	1407
ARTIFACT (A)	2164	SUBSTANCE ($)	1291
COGNITION (^)	1672	QUANTITY (Q)	1022

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. 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.

[7] Token-level measures consider both the supersense label and whether it begins or continues the mention.

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 Rodríguez, 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)*, 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-Aoussi, 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 Odjik, 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.