Supersense Tagging for Arabic: the MT-in-the-Middle Attack

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

Abstract

We consider the task of tagging Arabic nouns with WordNet supersenses. Three approaches are evaluated. The first uses an expert-crafted but limited-coverage lexicon, Arabic WordNet, and heuristics. The second uses unsupervised sequence modeling. The third and most successful approach uses machine translation to translate the Arabic into English, which is automatically tagged with English supersenses, the results of which are then projected back into Arabic. Analysis shows gains and remaining obstacles in four Wikipedia topical domains.

1 Introduction

A taxonomic view of lexical semantics groups word senses/usages into categories of varying granularities. WordNet *supersense tags* denote coarse semantic classes, including and (for nouns) and and (for verbs); these categories can be taken as the top level of a taxonomy. Nominal *supersense tagging* (Ciaramita and Johnson, 2003) is the task of identifying lexical chunks in the sentence for common as well as proper nouns, and labeling each with one of the 25 nominal supersense categories. Figure 1 illustrates two such labelings of an Arabic sentence. Like the narrower problem of named entity recognition, supersense tagging of text holds attraction as a way of inferring representations that move toward language independence. Here we consider the problem of nominal supersense tagging for Arabic, a language with ca. 300 million speakers and moderate linguistic resources, including a WordNet (Elkateb et al., 2006), annotated datasets (Maamouri et al., 2004; Hovy et al., 2006), monolingual corpora, and large amounts of Arabic-English parallel data.

The supervised learning approach that is used in state-of-the-art English supersense taggers (Ciaramita and Altun, 2006) is problematic for Arabic, since there are supersense annotations for only a small amount of Arabic text (65,000 words by Schneider et al., 2012, versus the 360,000 words that are annotated for English). Here, we reserve that corpus for development and evaluation, not training.

يتحكم مدير النوافذ في وضع وشكل نوافذ التطبيقات .

Ann-A	Gloss	Ann-B
	controls	
	manager	
	the-windows	
	in	
	configuration	
	and-layout	
	windows	
	the-applications	

'The window manager controls the configuration and layout of application windows.'

Figure 1: A sentence from the "X Window System" article with supersense taggings from two annotators and post hoc English glosses and translation.

We explore several approaches in this paper, the most effective of which is to (1) translate the Arabic sentence into English, returning the alignment structure between the source and target, (2) apply English supersense tagging to the target sentence, and (3) heuristically project the tags back to the Arabic sentence across these alignments. This "**MT-in-the-middle**" approach has also been successfully used for mention detection (Zitouni and Florian, 2008) and coreference resolution (Rahman and Ng, 2012).

We first discuss the task and relevant resources (§2), then the approaches we explored (§3), and finally present experimental results and analysis in §4.

2 Task and Resources

A gold standard corpus of sentences annotated with nominal supersenses (as in figure 1) facilitates automatic evaluation of supersense taggers. For development and evaluation we use

Proceedings of NAACL-HLT 2013, pages 661–667,
Atlanta, Georgia, 9–14 June 2013.

the AQMAR Arabic Wikipedia Supersense Corpus[1] (Schneider et al., 2012), which augmented a named entity corpus (Mohit et al., 2012) with nominal supersense tags. The corpus consists of 28 articles selected from four topical areas: **history** (e.g., "Islamic Golden Age"), **science** ("Atom"), **sports** ("Real Madrid"), and **technology** ("Linux"). Schneider et al. (2012) found the distributions of supersense categories in these four topical domains to be markedly different; e.g., most instances of (which includes kinds of software) occurred in the technology domain, whereas most s were found in the science domain.

The 18 test articles have 1,393 sentences (39,916 tokens) annotated at least once.[2] As the corpus was released with two annotators' (partially overlapping) taggings, rather than a single gold standard, we treat the output of each annotator as a separate test set. Both annotated some of every article; the first (**Ann-A**) annotated 759 sentences, the second (**Ann-B**) 811 sentences.

Lexicon. What became known as "supersense tags" arose from a high-level partitioning of synsets in the original English WordNet (Fellbaum, 1998) into *lexicographer files*. Arabic WordNet (AWN) (Elkateb et al., 2006) allows us to recover supersense categories for some 10,500 Arabic nominal types, since many of the synsets in AWN are cross-referenced to English WordNet, and can therefore be associated with supersense categories. Further, OntoNotes contains named entity annotations for Arabic (Hovy et al., 2006).

From these, we construct an Arabic supersense lexicon, mapping Arabic noun lemmas to supersense tags. This lexicon contains 23,000 types, of which 11,000 are multiword units. Token coverage of the test set is 18% (see table 1). Lexical units encountered in the test data were up to 9-ways supersense-ambiguous; the average ambiguity of in-vocabulary tokens was 2.0 supersenses.

Unlabeled Arabic text. For unsupervised learning we collected 100,000 words of Arabic Wikipedia text, not constrained by topic. The articles in this sample were subject to a minimum length threshold and are all cross-linked to corresponding articles in English, Chinese, and German.

Arabic→English machine translation. We used two independently developed Arabic-English MT systems. One (**QCRI**) is a phrase-based system (Koehn et al., 2003), similar to Moses (Koehn et al., 2007); the other (**cdec**) is a hierarchical phrase-based system (Chiang, 2007), as implemented in cdec (Dyer et al., 2010). Both were trained on about 370M tokens of parallel data provided by the LDC (by volume, mostly newswire and UN data). Each system includes preprocessing for Arabic morphological segmentation and orthographic normalization.[3] The **QCRI** system used a 5-gram modified Kneser-Ney language model that generated full-cased forms (Chen and Goodman, 1999). **cdec** used a 4-gram KN language model over lowercase forms and was recased in a post-processing step. Both language models were trained using the Gigaword v. 4 corpus. Both systems were tuned to optimize BLEU on a held-out development set (Papineni et al., 2002).

English supersense tagger. For English supersense tagging, an open-source reimplementation of the approach of Ciaramita and Altun (2006) was released by Michael Heilman.[4] This tagger was trained on the SemCor corpus (Miller et al., 1993) and achieves 77% F_1 in-domain.

3 Methods

We explored 3 approaches to the supersense tagging of Arabic: heuristic tagging with a lexicon, unsupervised sequence tagging, and MT-in-the-middle.

3.1 Heuristic Tagging with a Lexicon

Using the lexicon built from AWN and OntoNotes (see §2), our heuristic approach works as follows:

1. Stem and vocalize; we used MADA (Habash and Rambow, 2005; Roth et al., 2008).
2. Greedily detect word sequences matching lexicon entries from left to right.
3. If a lexicon entry has more than one associated supersense, Arabic WordNet synsets are

[1] http://www.ark.cs.cmu.edu/ArabicSST

[2] Our development/test split of the data follows Mohit et al. (2012), but we exclude two test set documents—"Light" and "Ibn Tolun Mosque"—due to preprocessing issues.

[3] **QCRI** accomplishes this using MADA (Habash and Rambow, 2005; Roth et al., 2008). **cdec** includes a custom CRF-based segmenter and standard normalization rules.

[4] http://www.ark.cs.cmu.edu/mheilman/questions

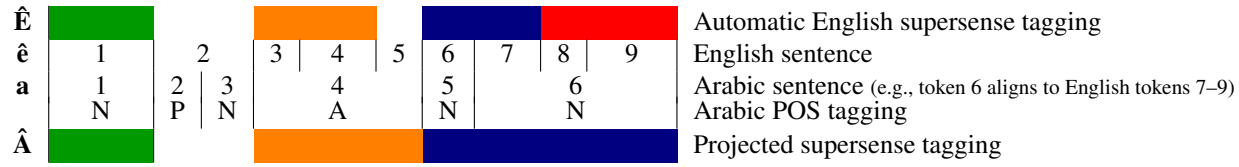

Figure 2: A hypothetical aligned sentence pair of 9 English words (with their supersense tags) and 6 Arabic words (with their POS tags). Step 4 of the projection procedure constructs the Arabic-to-English mapping $\{1\rightarrow \tfrac{1}{1}, 4\rightarrow \tfrac{4}{3}, \{5,6\}\rightarrow \tfrac{7}{6}\}$, resulting in the tagging shown in the bottom row.

weighted to favor earlier senses (presumed by lexicographers to be more frequent) and then the supersense with the greatest aggregate weight is selected. Formally: Let $senses(w)$ be the ordered list of AWN senses of lemma w. Let $senses(w, s) \subseteq senses(w)$ be those senses that map to a given supersense s. We choose $\arg\max_s(|senses(w, s)|/\min_{i:senses(w)_i \in senses(w,s)} i)$.

3.2 Unsupervised Sequence Models

Unsupervised sequence labeling is our second approach (Merialdo, 1994). Although it was largely developed for part-of-speech tagging, the hope is to use in-domain Arabic data (the unannotated Wikipedia corpus discussed in §2) to infer clusters that correlate well with supersense groupings. We applied the generative, feature-based model of Berg-Kirkpatrick et al. (2010), replicating a feature-set used previously for NER (Mohit et al., 2012)—including context tokens, character n-grams, and POS—and adding the vocalized stem and several stem shape features: 1) **ContainsDigit?**; 2) digits replaced by `#`; 3) digit sequences replaced by `#` (for stems mixing digits with other characters); 4) **YearLike?**—true for 4-digit numerals starting with `19` or `20`; 5) **LatinWord?**, per the morphological analysis; 6) the shape feature of Ciaramita and Altun (2006) (Latin words only). We used 50 iterations of learning (tuned on dev data). For evaluation, a many-to-one mapping from unsupervised clusters to supersense tags is greedily induced to maximize their correspondence on evaluation data.

3.3 MT-in-the-Middle

A standard approach to using supervised linguistic resources in a second language is **cross-lingual projection** (Yarowsky and Ngai, 2001; Yarowsky et al., 2001; Smith and Smith, 2004; Hwa et al., 2005; Mihalcea et al., 2007; Burkett and Klein, 2008; Burkett et al., 2010; Das and Petrov, 2011; Kim et al., 2012, who use parallel sentences extracted from Wikipedia for NER). The simplest such approach starts with an aligned parallel corpus, applies supersense tagging to the English side, and projects the labels through the word alignments. A supervised monolingual tagger is then trained on the projected labels. Preliminary experiments, however, showed that this underperformed even the simple heuristic baseline above (likely due to domain mismatch), so it was abandoned in favor of a technique that we call **MT-in-the-middle** projection.

This approach does not depend on having parallel data in the training domain, but rather on an Arabic→English machine translation system that can be applied to the sentences we wish to tag. The approach is inspired by token-level pseudo-parallel data methods of previous work (Zitouni and Florian, 2008; Rahman and Ng, 2012). MT output for this language pair is far from perfect—especially for Wikipedia text, which is distant from the domain of the translation system's training data—but, in the spirit of Church and Hovy (1993), we conjecture that it may still be useful. The method is as follows:

1. Preprocess the input Arabic sentence **a** to match the decoder's model of Arabic.
2. Translate the sentence, recovering not just the English output **ê** but also the derivation/alignment structure **z** relating words and/or phrases of the English output to words and/or phrases of the Arabic input.
3. Apply the English supersense tagger to the English translation, discarding any verbal supersense tags. Call the tagger output **Ê**.
4. Project the supersense tags back to the Arabic sentence, yielding **Â**: Each Arabic token $a \in$ **a** that is (a) a noun, or (b) an adjective following 0 or more adjectives following a noun is mapped to the first English supersense mention in **Ê** containing some word aligned to a. Then, for each English supersense men-

		Coverage			Ann-A			Ann-B		
		Nouns	All Tokens	Mentions	P	R	F_1	P	R	F_1
Lexicon heuristics (§3.1)		8,058 33%	8,465 18%	8,407	32 55	16 29	21.6 37.9	29 53	15 27	19.4 35.6
Unsupervised (§3.2)					20 59	16 48	17.5 52.6	14 56	10 39	11.6 45.9
MT-in-the-middle (§3.3)	**QCRI**	14,401 59%	16,461 35%	12,861	34 65	27 50	29.9 56.4	36 64	28 51	31.6 56.6
	cdec	14,270 58%	15,542 33%	13,704	**37 69**	**31 57**	**33.8 62.4**	**38 67**	**32 56**	**34.6 61.0**
MTitM + Lex.	**cdec**	16,798 68%	18,461 40%	16,623	**35 64**	**36 65**	**35.5 64.6**	**36 63**	**36 63**	**36.0 63.2**

Table 1: Supersense tagging results on the test set: coverage measures[5] and gold-standard evaluation—exact labeled/unlabeled[6] mention precision, recall, and F-score against each annotator. The last row is a hybrid: MT-in-the-middle followed by lexicon heuristics to improve recall. Best single-technique and best hybrid results are bolded.

tion, all its mapped Arabic words are grouped into a single mention and the supersense category for that mention is projected. Figure 2 illustrates this procedure. The **cdec** system provides word alignments for its translations derived from the training data; whereas **QCRI** only produces phrase-level alignments, so for every aligned phrase pair $\langle \bar{a}, \bar{e} \rangle \in \mathbf{z}$, we consider every word in $\bar{a}$ as aligned to every word in $\bar{e}$ (introducing noise when English supersense mention boundaries do not line up with phrase boundaries).

4 Experiments and Analysis

Table 1 compares the techniques (§3) for full Arabic supersense tagging.[7] The number of nouns, tokens, and mentions covered by the automatic tagging is reported, as is the mention-level evaluation against human annotations. The evaluation is reported separately for the two annotators in the dataset.

With heuristic lexicon lookup, 18% of the tokens are marked as part of a nominal supersense mention. Both labeled and unlabeled mention recall with this method are below 30%; labeled precision is about 30%, and unlabeled mention precision is above 50%. From this we conclude that the biggest problems are (a) out-of-vocabulary items and (b) poor semantic disambiguation of in-vocabulary items.

The unsupervised sequence tagger does even worse on the labeled evaluation. It has some success at *detecting* supersense mentions—unlabeled recall is substantially improved, and unlabeled precision is slightly improved. But it seems to be much worse at assigning semantic categories; the number of labeled true positive mentions is actually *lower* than with the lexicon-based approach.

MT-in-the-middle is by far the most successful single approach: both systems outperform the lexicon-only baseline by about 10 F_1 points, despite many errors in the automatic translation, English tagging, and projection, as well as underlying linguistic differences between English and Arabic. The baseline's unlabeled recall is doubled, indicating substantially more nominal expressions are detected, in addition to the improved labeled scores.

We further tested simple **hybrids** combining the lexicon-based and MT-based approaches. Applying MT-in-the-middle first, then expanding token coverage with the lexicon improves recall at a small cost to precision (table 1, last row). Combining the techniques in the reverse order is slightly worse than MT-based projection without consulting the lexicon.

MT-in-the middle improves upon the lexicon-only baseline, yet performance is still dwarfed by the supervised English tagger (at least in the SemCor evaluation; see §2), and also well below the 70% inter-annotator F_1 reported by Schneider et al. (2012). We therefore examine the weaknesses of our approach for Arabic.

4.1 MT for Projection

In analyzing our projection framework, we performed a small-scale MT evaluation with the Wikipedia data. Reference English translations for 140 Arabic Wikipedia sentences—5 per article in the corpus—were elicited from a bilingual linguist. Table 2 compares the two systems under three standard metrics of overall sentence translation quality.[8]

[5] The unsupervised evaluation greedily maps clusters to tags, separately for each version of the test set; coverage numbers thus differ and are not shown here.

[6] Unlabeled tagging refers to noun chunk *detection* only.

[7] It was produced in part using the `chunkeval.py` script: see `https://github.com/nschneid/pyutil`

[8] BLEU (Papineni et al., 2002); METEOR (Banerjee and Lavie, 2005; Lavie and Denkowski, 2009), with default options;

تتكون الذرة من سحابة من الشحنات السالبة (الإلكترونات) تحوم حول نواة موجبة الشحنة صغيرة جدا في الوسط .

QCRI: corn consists of a negative shipments (electron hovering around the nucleus of the shipment) are very small in the center . (3/6)
cdec: The corn is composed of negative shipments (electronics) cloud hovering over the nucleus of a very small positive shipment in the center . (2/6)

تأهلت البرتغال للنهائيات بمتنهي السهولة ، فالبرتغال لم تهزم طوال مشوار التصفيات .

QCRI: Portugal qualified for the finals very easily , Portugal defeated throughout the mission liquidations . (3/5)
cdec: Portugal qualified easily for the finals , Portugal unbeaten throughout the journey . (3/4)

Figure 3: Example Arabic inputs and the outputs of the two MT systems, with lexical projection precision ratios.

While the resulting number of sentences with references is far from ideal and there is only one reference translation for each, all three measures favor **QCRI**.

For a targeted measure of *lexical* translation quality of noun expressions, we elicited acceptability judgments from a bilingual annotator for translated and supersense-projected phrases. Given each MT system output (for the same 140 sentences) in which mentions predicted by the English supervised tagger were highlighted, along with the Arabic source sentence, the judge was asked for each English mention whether it was a valid translation.[9] We call this **lexical projection precision**. Figure 3 shows examples, and the last column of table 2 gives overall statistics. Upwards of 90% of the lexical translations were accepted; as with the automatic MT measures, **QCRI** slightly outperforms **cdec** (especially in the technology and sports domains[10]). Of the problematic lexical translations, some are almost certainly domain effects: e.g., *corn* or *maize* instead of *atom*. Others are more nuanced, e.g., *shipments* for *charges* and *electronics* for *electrons*. Transliteration errors included *IMAX* in place of *EMACS* and *genoa lynx* for *GNU Linux*. However, lexical projection precision seems to be a relatively small part of the problem, especially considering that not all translation errors lead to supersense tagging errors.

Lexical projection *recall* was not measured, but noun token coverage (see table 1) is instructive. Most noun tokens ought to be tagged; 77% is the noun coverage rate in the gold standard. In table 1, noun coverage gains track overall improvements.

	BLEU	METEOR	TER	Lex. Prec.
QCRI	**32.86**	**32.10**	**0.46**	**91.9%**
cdec	28.84	31.38	0.49	90.0%

Table 2: MT quality measures comparing the two systems over 140 sentences. The first three are automatic measures with 1 reference translation. For the fourth, a bilingual judged the translation acceptability of phrases that were identified as supersense mentions by the English tagger (**lexical projection precision**).

If **QCRI** produces better translations, why is **cdec** more useful for supersense tagging? As noted in §3.3, **cdec** gives word-level alignments (even when the decoder uses large phrasal units for translation), allowing for more precise projections.[11] We suspect this is especially important in light of findings that larger phrase sizes are indicative of better translations (Gamon et al., 2005), so these are exactly the cases where we expect the translations to be valuable.

5 Conclusion

To our knowledge, this is the first study of automatic Arabic supersense tagging. We have shown empirically that an MT-in-the-middle technique is most effective of several approaches that do not require labeled training data. Analysis sheds light on several challenges that would need to be overcome for better Arabic lexical semantic tagging.

Acknowledgments

We thank Wajdi Zaghouani for the translation of the Arabic Wikipedia MT set, Francisco Guzman and Preslav Nakov for the output of QCRI's MT system, and Waleed Ammar 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.

and translation edit rate (TER) (Snover et al., 2006)

[9] The judge did not see alignments or supersense categories.

[10] For technology articles, the differences in F_1 scores between the two systems were 6.1 and 4.2 for **Ann-A** and **Ann-B**, respectively. For sports the respective differences were 4.3 and 4.4. In the other domains the differences never exceeded 3.3. Interestingly, this is the only generalization about topical domain performance we were able to find that holds across annotators and systems, in contrast with the stark topical effects observed by Mohit et al. (2012) for NER.

[11] Our experiments use **QCRI** as an off-the-shelf system. As a reviewer notes, it could be retrained to produce word-level alignments, which would likely improve the accuracy of supersense tag projection.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010)*, pages 582–590, Los Angeles, California, June. Association for Computational Linguistics.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 877–886, Honolulu, Hawaii, October. Association for Computational Linguistics.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010. Learning better monolingual models with unannotated bilingual text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010)*, pages 46–54, Uppsala, Sweden, July. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, October.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, June.

Kenneth W. Church and Eduard H. Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258, December.

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.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 600–609, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7–12, Uppsala, Sweden, July. Association for Computational Linguistics.

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, May.

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

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th European Association for Machine Translation Conference (EAMT 2005)*, pages 103–111, Budapest, May.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan, June. Association for Computational Linguistics.

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 2006)*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, September.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from Wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 694–702, Jeju Island, Korea, July. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceed-*

ings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003), pages 48–54, Edmonton, Canada. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Alon Lavie and Michael J. Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115, September.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: building a large-scale annotated Arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 976–983, Prague, Czech Republic, June. 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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)*, pages 720–730, Montréal, Canada, June. Association for Computational Linguistics.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08: HLT*, pages 117–120, Columbus, Ohio, June. Association for Computational Linguistics.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse lexical semantic annotation with supersenses: an Arabic case study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 253–258, Jeju Island, Korea, July. Association for Computational Linguistics.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: using English to parse Korean. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, pages 49–56, Barcelona, Spain.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, USA, August.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'01)*, Pittsburgh, Pennsylvania, USA, June.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT'01)*, San Diego, California, USA, March.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 600–609, Honolulu, Hawaii, October. Association for Computational Linguistics.