SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM)

Nathan Schneider
School of Informatics
University of Edinburgh
Edinburgh, UK

Dirk Hovy **Anders Johannsen**
Center for Language Technology
University of Copenhagen
Copenhagen, Denmark

Marine Carpuat
Computer Science Dept.
University of Maryland
College Park, Maryland, USA

nschneid@inf.ed.ac.uk, krx628@hum.ku.dk, anders@johannsen.com, marine@cs.umd.edu

Abstract

This task combines the labeling of multiword expressions and supersenses (coarse-grained classes) in an explicit, yet broad-coverage paradigm for lexical semantics. Nine systems participated; the best scored 57.7% F_1 in a multi-domain evaluation setting, indicating that the task remains largely unresolved. An error analysis reveals that a large number of instances in the data set are either hard cases, which no systems get right, or easy cases, which all systems correctly solve.

1 Introduction

Grammatical analysis tasks, e.g., part-of-speech tagging, are rather successful applications of natural language processing (NLP). They are *comprehensive*, i.e., they operate under the assumption that all grammatically-relevant parts of a sentence will be analyzed: We do not expect a POS tagger to only know a subset of the tags in the language. Most POS tags accommodate unseen words and adapt readily to new text genres. Together, these factors indicate a representation which achieves *broad coverage*.

Explicit analysis of lexical semantics, by contrast, has been more difficult to scale to broad coverage owing to limited comprehensiveness and extensibility. The dominant paradigm of fine-grained word sense disambiguation, WordNet (Fellbaum, 1998), is difficult to annotate in corpora, results in considerable data sparseness, and does not readily generalize to out-of-vocabulary words. While the main corpus with WordNet senses, SemCor (Miller et al., 1993), does reflect several text genres, it is hard to expand SemCor-style annotations to new genres, such as social web text or transcribed speech. This severely limits the applicability of SemCor-based NLP tools and restricts opportunities for linguistic studies of lexical semantics in corpora.

To address this limitation, in the DiMSUM 2016 shared task,[1] we challenged participants to analyze the lexical semantics of English sentences with a tagset integrating **multiword expressions** and **noun and verb supersenses** (following Schneider and Smith, 2015), on multiple nontraditional genres of text. By moving away from fine-grained sense inventories and lexicalized, language-specific[2] annotation, we take a step in the direction of broad-coverage, coarse-grained lexical semantic analysis. We believe this departure from the classical lexical semantics paradigm will ultimately prove fruitful for a variety of NLP applications in a variety of genres.

The integrated lexical semantic representation (§2, §3) has been annotated in an extensive benchmark data set comprising several nontraditional domains (§4). Objective, controlled evaluation procedures (§5) facilitate a comparison of the 9 systems submitted as part of the official task (§6). While the systems range in performance, all are below 60% in our composite evaluation, suggesting that further work is needed to make progress on this difficult task.

2 Background

Multiword expressions. Most contemporary approaches to English syntactic and semantic analysis treat space-separated words as the basic units of structure. However, this fails to reflect the basic units of meaning for sentences with noncompositional or idiosyncratic expressions, such as:

(1) The staff leaves a lot to be desired .

(2) I googled restaurants in the area and Fuji Sushi came up and reviews were great so I made a carry out order of : L 17 .

[1] http://dimsum16.github.io/

[2] Though our data set is limited to English, the representation is applicable to other languages: see §2.

In these sentences, *a lot*, *leaves... to be desired*, *Fuji Sushi*, *came up*, *made... order*, and *carry out* are all **multiword expressions** (MWEs): their combined meanings can be thought of as “prepackaged” in a single lexical expression that happens to be written with spaces. MWEs such as these have attracted a great deal of attention within computational semantics; see Baldwin and Kim (2010) for a review. Schneider et al. (2014b) introduced an English corpus resource annotated for heterogenous MWEs, suitable for training and evaluating general-purpose MWE identification systems (Schneider et al., 2014a). Prior to that, most MWE evaluations were focused on particular constructions such as noun compounds (recently: Constant and Sigogne, 2011; Green et al., 2012; Ramisch et al., 2012; Vincze et al., 2013), though the corpus and identification system of Vincze et al. (2011) targets several kinds of MWEs.

Importantly, the MWEs in Schneider et al.’s (2014b) corpus are not required to be contiguous, but may contain **gaps** (viz.: *made... order*). The corpus also contains qualitative labels indicating the strength of MWEs, either **strong** (mostly non-compositional) or **weak** (compositional but idiomatic). For simplicity we only include strong MWEs in this task.

Supersenses. As noted above, relying on WordNet-like fine-grained, lexicalized **senses** creates problems for annotating at a large scale and covering new domains and languages. Named entity recognition (NER) does not suffer from these problems, as it uses a much smaller number of coarse-grained classes. However, these classes only apply to a subset of the nouns in a sentence and exclude verbs and adjectives. They therefore provide far from complete coverage in a corpus.

Noun and verb **supersenses** (Ciaramita and Altun, 2006) offer a middle ground in granularity: they generalize named entity classes to cover all nouns (with 26 classes), but also cover verbs (15 classes)—see table 1—and provide a human-interpretable high-level clustering. WordNet supersenses for adjectives and adverbs nominally exist, but are based on morphosyntactic rather than semantic properties. There is, however, recent work on developing supersense taxonomies for English adjectives and prepositions (Tsvetkov et al., 2014; Schneider et al., 2015).

N:TOPS	N:OBJECT	V:COGNITION
N:ACT	N:PERSON	V:COMMUNICATION
N:ANIMAL	N:PHENOMENON	V:COMPETITION
N:ARTIFACT	N:PLANT	V:CONSUMPTION
N:ATTRIBUTE	N:POSSESSION	V:CONTACT
N:BODY	N:PROCESS	V:CREATION
N:COGNITION	N:QUANTITY	V:EMOTION
N:COMMUNICATION	N:RELATION	V:MOTION
N:EVENT	N:SHAPE	V:PERCEPTION
N:FEELING	N:STATE	V:POSSESSION
N:FOOD	N:SUBSTANCE	V:SOCIAL
N:GROUP	N:TIME	V:STATIVE
N:LOCATION	V:BODY	V:WEATHER
N:MOTIVE	V:CHANGE	

Table 1: The 41 noun and verb supersenses in WordNet.

The inventory for nouns and verbs originates from the top-level organization of WordNet, but can be applied directly to annotate new data—including out-of-vocabulary words in English or other languages (Schneider et al., 2012; Johannsen et al., 2014). Similar to NER, supersense tagging approaches have generally used statistical sequence models and have been evaluated in English, Italian, Chinese, Arabic, and Danish.[3]

Features based on supersenses have been exploited in downstream semantics tasks such as preposition sense disambiguation, noun compound interpretation, question generation, and metaphor detection (Ye and Baldwin, 2007; Hovy et al., 2010; Tratz and Hovy, 2010; Heilman, 2011; Hovy et al., 2013; Tsvetkov et al., 2013).

Relationship between MWEs and supersenses. We believe that MWEs and supersenses should be tightly coupled: idiomatic combinations such as MWEs are best labeled holistically, since their joint supersense category will often differ from that of the individual words. For example, *spill the beans* in its literal interpretation would receive supersenses V:CONTACT and N:FOOD, whereas the idiomatic interpretation, ‘divulge a secret’, is represented as an MWE holistically tagged as V:COMMUNICATION. Schneider and Smith (2015) develop this idea at

[3] Evaluations used English SemCor (Ciaramita and Altun, 2006; Paaß and Reichartz, 2009), English-Italian MultiSemCor (Picca et al., 2008, 2009; Attardi et al., 2010), the Italian Syntactic-Semantic Treebank and Italian Wikipedia (Attardi et al., 2010; Rossi et al., 2013), Chinese Cilin (Qiu et al., 2011), Arabic Wikipedia (Schneider et al., 2013), and the Danish CLARIN Reference Corpus (Martínez Alonso et al., 2015).

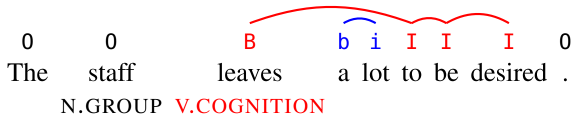


Figure 1: Illustration of the target representation. MWE positional markers are shown above the sentence and noun and verb supersenses below the sentence. Links illustrate the behavior of the MWE tags. The supersense labeling must respect the MWEs; thus, V.COGNITION applies to a four-word unit—*to*, *be*, and *desired* must not receive separate supersenses from *leaves*.

length, and provide a web reviews data set with the integrated annotation. Here, we expand the paradigm to additional domains and compare the performance of several systems.

3 Representation

The analysis for each sentence is represented as a sequence of paired MWE and supersense tags. Figure 1 illustrates the MWE part above the sentence and the supersense part below the sentence.

The MWE portion is a BIO-style (Ramshaw and Marcus, 1995) positional marker. Of the schemes discussed by Schneider et al. (2014a), we adopt the 6-tag scheme, which uses case to allow gaps in an MWE (lowercase tag variants mark tokens within a gap). The positions are thus `O`, `o`, `B`, `b`, `I`, `i`. Systems are expected to ensure that the full tag sequence for a sentence is valid: global validity can be enforced with first-order constraints to prohibit invalid bigrams such as `O I` and `b I` (see Schneider et al., 2014a for details).

Because strong MWEs receive a supersense as a unit (if at all), `I` and `i` are never accompanied by a supersense label. `O` or `o` indicates that the token is not part of any MWE, but many such tokens do bear a noun or verb supersense.

This task uses a CoNLL-style main **file format** consisting of one line per token, each line having 9 tab-delimited columns. Scripts to convert to and from the `.sst` format, which displays one sentence per line and contains annotations in a JSON data structure, are provided as well.

4 Data

The task built upon two existing data sets of social web text, which were harmonized to form the training data. Four new samples from three domains were newly annotated to form the test set. The train and test sets are summarized in tables 2 and 3 and are publicly available on the web.[4]

The domains covered are **online customer reviews**, **tweets**, and **TED talks**. This section describes, for each domain, how its component data sets were sampled, preprocessed, and annotated.

4.1 Annotation Process

We compiled data sets from various sources, with varying degrees of existing pre-annotation. Unless already provided, we added Universal POS tags as defined by the Universal Dependencies project (Nivre et al., 2015), and baseline supersenses (heuristically using the most frequent WordNet sense, and in some cases grouping sequences of proper nouns as MWEs). The pre-annotated supersenses were then manually corrected by a trained annotator, who simultaneously annotated the sentence for comprehensive MWEs.

The annotator (a linguist) was trained by the first author of this paper using Schneider and Smith's (2015) web interface and annotation guidelines. Prior to starting on the data sets for this task, the annotator devoted approximately 8 hours to training practice on a separate data set which already had a gold standard. Periodic feedback was given on initial annotations as the annotator grew accustomed to the conventions. The annotator spent approximately 50 hours on DiMSUM data (not including the initial training phase), which amounts to roughly 90 seconds per sentence.

In order to estimate inter-annotator agreement (IAA), the first author independently annotated a sample of Ritter tweets (§4.3) in 6 groups of 11 sentences, spaced out across the main annotator's annotation batches. IAA estimates for these sets ranged from 60% to 75% F_1 for MWEs, and 67%–80% accuracy for supersenses (on tokens which had supersenses in both annotations). Resources did not allow for more systematic double annotation and IAA estimation throughout the data.

The test set newly annotated for this task comprises exactly 1,000 sentences and exactly 16,500 words. 3,120 word tokens (19%) differ from the pre-annotation with respect to gold MWE boundaries

[4] `https://github.com/dimsum16/dimsum-data`

	Domain	Source corpus	UPOS (UD 1.2–style)	Docs	Sents	Words	w/s	#lemmas
Train	REVIEWS	STREUSLE 2.1 (Schneider and Smith, 2015)	Conv. from PTB parses	723	3,812	55,579	14.6	5,052
	TWEETS	Lowlands (tweets w/ URLs) (Johannsen et al., 2014)	Conv. from Petrov-style	N/A	200	3,062	15.3	1,201
	TWEETS	Ritter (Ritter et al., 2011; Johannsen et al., 2014)	Conv. from Petrov-style	N/A	787	15,185	19.3	3,819
			Train Total		4,799	73,826	15.4	7,988
Test	REVIEWS	Trustpilot (Hovy and Søgaard, 2015)	Conv. from Petrov-style	N.A.	340	6,357	18.7	1,365
	TWEETS	Tweebank (Kong et al., 2014)	Conv. from TweetNLP POS in FUDG parses	N/A	500	6,627	13.3	1,786
	TED	NAIST-NTT (⊂ IWSLT train) (Cettolo et al., 2012; Neubig et al., 2014)	Conv. from PTB parses	10	100	2,187	21.9	630
	TED	IWSLT test (Cettolo et al., 2012)	Auto	6	60	1,329	22.2	457
			Test Total		1,000	16,500	16.5	3,160
			REVIEWS Total		4,152	61,936	14.9	5,477
			TWEETS Total		1,487	24,874	16.7	5,464
			TED Total		160	3,516	22.0	900
			Grand Total		5,799	90,326	15.6	9,321

Table 2: Source datasets and preprocessing to obtain 17-tag Universal POS tags (UPOS) version 1.2. Most sources already contained some form of POS tags, which we automatically converted to UPOS. We added missing necessary distinctions—e.g., UD-style UPOS distinguishes auxiliaries from main verbs, but Petrov-style (Petrov et al., 2011), PTB (Marcus et al., 1993), or TweetNLP (Owoputi et al., 2013) POS tagsets do not. Disambiguation was done manually or via a gold parse, where available. We also modified the tokenization of the Tweebank data, to be consistent with UPOS conventions for English (e.g., separating clitics).

Only some portions of the data group sentences into documents: N/A = not applicable; N.A. = not available.

						Gappy		% tokens	N SS		V SS	
	Domain	**Source corpus**	**MWEs+Supersenses**	**Words**	**MWEs**	**MWEs**		**in MWE**	**units**	**MWE**	**units**	**MWE**
Train	REVIEWS	STREUSLE	Gold	55,579	3,117	397	13%	13%	9,112	13%	7,689	13%
	TWEETS	Lowlands	Gold—revised	3,062	276	5	2%	22%	741	31%	281	7%
	TWEETS	Ritter	Gold—revised	15,185	839	65	8%	13%	2,738	20%	1,893	10%
			Train Total	73,826	4,232	467	11%	13%	12,591	16%	9,863	12%
Test	REVIEWS	Trustpilot	Gold—new	6,357	327	13	4%	12%	1,055	23%	848	6%
	TWEETS	Tweebank	Gold—new	6,627	362	20	6%	13%	899	21%	911	8%
	TED	NAIST-NTT	Gold—new	2,187	93	2	2%	9%	373	16%	278	7%
	TED	IWSLT test	Gold—new	1,329	55	1	2%	9%	228	15%	153	3%
			Test Total	16,500	837	36	4%	12%	2,555	21%	2,190	7%
			REVIEWS Total	61,936	3,444	410	12%	13%	10,167	14%	8,537	12%
			TWEETS Total	24,874	1,477	90	6%	14%	4,378	22%	3,085	9%
			TED Total	3,516	148	3	2%	9%	601	16%	431	6%
			Grand Total	90,326	5,069	503	10%	13%	15,146	17%	12,053	11%

Table 3: Annotated datasets: status of lexical semantic annotations (retained, revised, or newly annotated for this task) per subcorpus; word token and MWE instance counts; number and proportion (out of all MWEs) that are gappy; proportion of tokens that belong to an MWE; number of units labeled with a noun supersense, and proportion that are MWEs; likewise for verb supersenses.

Additional statistics relatively consistent across domains: MWEs per word: mean/median .055 (lowest: TED, .044; highest: STREUSLE, .090). Supersenses per word: mean/median 0.3. Just 8 MWEs contain more than one gap (all in STREUSLE or Ritter).

and/or supersenses.[5] In addition, portions of the training data were reannotated for improved quality and consistency with the STREUSLE annotations, as explained below.

4.2 REVIEWS

Training. The REVIEWS part of the training data consists of the STREUSLE corpus (Schneider et al., 2014b; Schneider and Smith, 2015),[6] comprising comprehensive multiword expression and supersense annotations on a 55,000-token portion of the English Web Treebank (EWTB; Bies et al., 2012) made up of 723 online user reviews for services (such as restaurants and beauticians).

STREUSLE annotation was done by linguists, who took pains to establish conventions and resolve disagreements. Each sentence was annotated independently by at least 2 annotators; disagreements were resolved by negotiation.

The task release is based on version 2.1 of STREUSLE, with weak MWEs removed and Penn Treebank–style POS tags replaced with Universal POS tags.[7]

Test. The test portion comprises 340 sentences (6,357 tokens) from the online review site Trustpilot, a subset of the data used in Hovy and Søgaard (2015) (the website as a general resource was described in Hovy et al. (2015)). The reviews were chosen to obtain a demographic balance (by age, gender, and location), and contained gold POS tags.

4.3 TWEETS

Training. Johannsen et al. (2014) recently annotated two samples of 987 Twitter messages (18,000 words) with supersenses: (a) the POS+NER-annotated data set of Ritter et al. (2011), and (b) Plank et al.'s (2014) sample of 200 tweets.[8] Annotators were shown pre-annotations from a heuristic supersense chunking/tagging system (based on the most frequent sense of each word) and asked to correct the boundaries and supersense labels. Though there was no explicit MWE annotation phase, many of the multiword chunks tagged with a noun or verb supersense would be considered MWEs.

We fully reannotated both data sets to match the conventions of the REVIEWS data from the STREUSLE corpus. The annotator examined every sentence and corrected any MWE or supersense decisions deemed to be inconsistent with the guidelines.

Test. Our test set consists of 500 tweets (6,627 tokens) taken from the Tweebank corpus (Kong et al., 2014),[9] which already contained some gold-standard MWEs. We converted the POS tags from gold ARK TweetNLP POS + FUDG dependencies to UPOS and had an annotator supply supersenses.

4.4 TED TALKS

Test. To test the broad-coverage aspect of the submitted systems, the test set contained a "surprise" domain. We opted to sample transcribed sentences from TED talks. Because individual TED talks tend to heavily repeat vocabulary, we took the first 10 sentences from each of 16 documents in order to achieve a lexically diverse sample. Specifically, we chose (a) 100 sentences (2,187 tokens) from the 10 talks in the NAIST-NTT Ted Talk Treebank[10] (Neubig et al., 2014) (which in turn is a subset of the IWSLT training data); and (b) 60 sentences (1,329 tokens) from the IWSLT test data (Cettolo et al., 2012).[11] The latter 6 documents were chosen to maximize language pair diversity.[12]

We induced parts of speech by conversion from the gold PTB trees for the NAIST-NTT data, and

[5] On the surface, this might be taken to mean that the accuracy of the heuristic baseline used for pre-annotation is 81%. However, because the annotator saw the pre-annotation, we expect that this agreement rate is higher than if the gold standard had been produced from scratch.

[6] `http://www.ark.cs.cmu.edu/LexSem/`

[7] The PTB-to-UPOS conversion script is available at: `http://tiny.cc/ptb2upos`

[8] The supersense-annotated tweets are available at `https://github.com/coastalcph/supersense-data-twitter`.

[9] `http://www.cs.cmu.edu/~ark/TweetNLP/`

[10] `http://ahclab.naist.jp/resource/tedtreebank/`

[11] `https://wit3.fbk.eu/`

[12] These 6 talks are known to have been translated from English into (at least) the following languages: {ar, de, es, fa, he, hi, it, ko, nl, th, vi, zh}. Additionally, we note that 4 of the documents have Czech (cs) translations, while the other 2 have French (fr) translations.

Neubig et al. (2014) report that all the 10 documents in the NAIST-NTT Treebank have been translated from English into the following 18 languages: {ar, bg, de, el, es, fr, he, it, ja, ko, nl, pl, pt-BR, ro, ru, tr, zh-CN, zh-TW}. Many additional languages are represented for subsets of the documents.

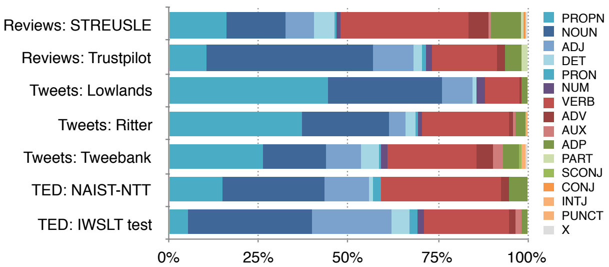


Figure 2: Counts of MWE occurrences, grouped by the POS of the first word in the MWE. Blue bars represent POSes that tend to start nominal MWEs; red bars roughly capture verbal MWEs.

for the remaining data, by automatic tagging with an averaged structured perceptron model (Rungsted[13]) trained on the English Universal Dependencies v1.2 treebank (Nivre et al., 2015).[14]

4.5 Comparing Domains

A natural question to ask about lexical semantic annotations is whether we observe strong differences between domains. For example, which kinds of multiword expressions and which kinds of supersenses occur more often in some domains than in others? In this section, we report our observations but do not make any strong claims about their generality, for the following reasons: the samples are not necessarily representative of their domains overall, and, in fact, may have been sampled in a biased way (e.g., the Lowlands sample was limited to tweets containing a URL, and as a result, most of these tweets are headlines and advertisements). Furthermore, the annotation procedures differed by subcorpus, likely biasing the results.

MWEs. Figure 2 summarizes MWEs in the seven subcorpora with respect to syntactic status. Colors represent the POS tag of the first word in the MWE. Starting with proper nouns, the blue bars indicate POS tags that tend to begin nominal MWEs (noun, adjective, determiner, etc.). Red bar POS tags are characteristic of verbal MWEs. The remaining bars are prepositional (dark green) and other miscellaneous tags, which collectively comprise no more than 10% of the MWEs in each subcorpus.

It is worth noting that in this plot, subcorpora within the same domain are sometimes more divergent than subcorpora in different domains. Lowlands stands out as having a large share of proper noun MWEs—presumably due to the headline-oriented nature of the sample. STREUSLE has the smallest proportion of nominal MWEs, perhaps owing to the way it was annotated: initial rounds of STREUSLE annotation targeted MWEs only, with noun and verb supersenses added only later; whereas in the other data sets, MWE and supersense annotation were performed jointly, so annotator attention may have been focused on nominal and verbal expressions rather than other MWEs.

Supersenses. In the spirit of Schneider et al. (2012), we performed an analysis to see which supersenses were more characteristic of some domains than others. Figure 3 plots the relative frequency (out of all supersense-labeled units) of each supersense in each of the three domains. We use the REVIEWS domain as base frequency: relative to that, the x-axis is the supersense's occurrence rate in the TWEETS domain, and the y-axis represents the rate for the TED talks.

These plots show some clear outliers: among nouns (left plot), N.GROUP and N.FOOD are overrepresented in REVIEWS relative to the other domains—unsurprising because restaurants and other businesses are prominent in this subcorpus. On the other hand, N.PERSON is underrepresented in REVIEWS. N.TIME and N.COMMUNICATION are more popular in the TWEETS domain than the others. Among verbs (right plot), V.STATIVE is underrepresented, apparently due to the relative rarity of the copula (which often can be safely omitted in headlines and other telegraphic messages without obscuring the meaning).

5 Evaluation

Submission conditions. We invited submissions in multiple data conditions. The **open** condition encouraged participants to make wide use of any and all available resources, including for distant or direct supervision. A **closed** condition encouraged controlled comparisons of algorithms by limiting their training to specific resources distributed for the task. Lastly, we allowed for a **semi-supervised closed** condition, in which use of a specific large unla-

[13] `https://github.com/coastalcph/rungsted`

[14] `http://hdl.handle.net/11234/1-1548`

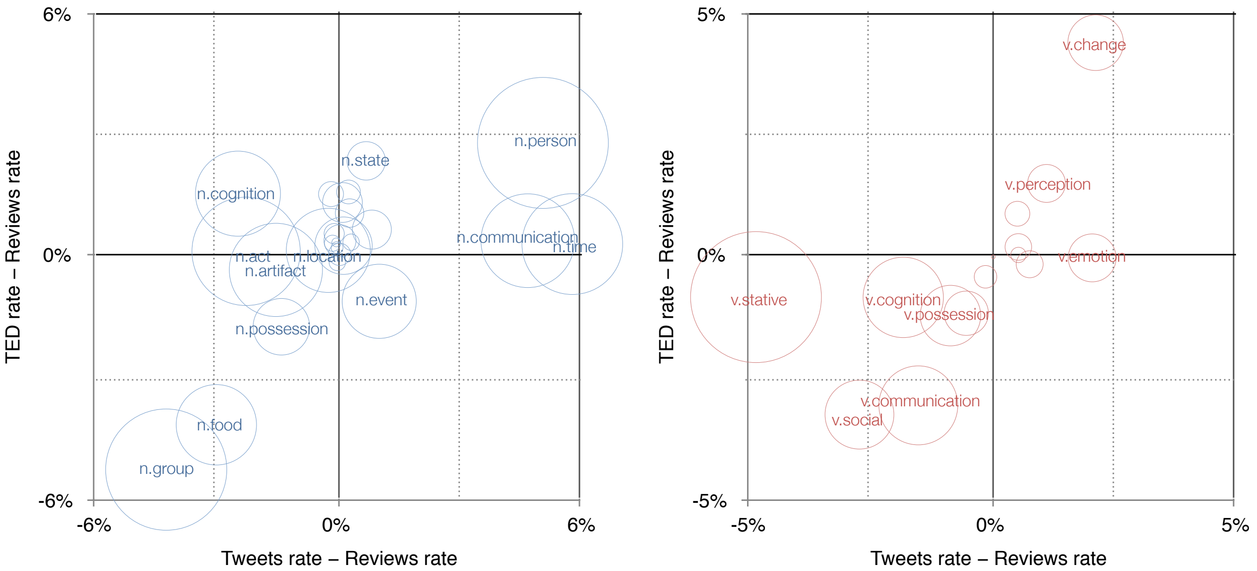


Figure 3: Supersense rate differences by domains, compared to reviews data set. Circle area proportional to the supersense's *total* frequency across all domains. Noun supersenses on the left, verb supersenses on the right. Each domain's rate is microaveraged across its subcorpora; thus, larger subcorpora weigh more heavily than smaller subcorpora in the same domain.

beled corpus—the Yelp Academic Dataset[15]—was permitted. Teams were permitted to submit no more than one run per condition. Only one team submitted a system in the semi-supervised closed condition.

All conditions had access to: 1) the annotated data we provided; 2) Brown clusterings (Brown et al., 1992) computed from large corpora of tweets and web reviews;[16] and 3) the English WordNet lexicon. The input at test time included POS tags.

No sentence-level metadata was provided in the input at test time: test set sentence IDs were obscured to hide the source domain, and the order of sentences was randomized to remove document structure. The training data, however, marked the domain from which the sentence was drawn (REVIEWS or TWEETS); systems were free to make use of this information, so long as it was not required as part of the input at test time.

Scoring. We provided an evaluation script to allow participants to check the format of system output and to compute all official scores.

The **MWE** measure looks at precision, recall, and F_1 of the identified MWEs. Tokens not involved in a predicted or gold MWE do not factor into this measure. To award partial credit for partial overlap between a predicted MWE and a gold MWE, these scores are computed based on *links* between consecutive tokens in an expression (Schneider et al., 2014a). The tokens must appear in order but do not need to be adjacent. The precision is the proportion of predicted links whose words both belong to the same expression in the gold standard. Recall is the same as precision, but swapping the predicted and gold annotations.[17] Figure 4 defines this measure in detail and illustrates the calculations for an example.

To isolate the **supersense** classification performance, we compute precision, recall, and F_1 of the supersense-labeled word tokens. The numerator of both precision and recall is the number of tokens labeled with the correct supersense. (This interacts slightly with MWE identification, however, as supersenses are only marked on the first token of MWEs. We do not mark supersenses on all words of the MWE to avoid giving MWEs a disproportionate influence on the supersense score.)

Finally, **combined** precision, recall, and F_1 aggregate the MWE and supersense subscores. The combined precision ratio is computed from the MWE

[15] `https://www.yelp.com/academic_dataset`

[16] I.e., TweetNLP clusters (`http://www.cs.cmu.edu/~ark/TweetNLP/`) and the Yelp Academic Dataset clusters used in AMALGrAM (`http://www.cs.cmu.edu/~ark/LexSem/`).

[17] This computation on the basis of links is a slight simplification of the MUC coreference measure (Vilain et al., 1995).

MWE Precision: The proportion of predicted links whose words both belong to the same expression in the gold standard.
MWE Recall: Same as precision, but swapping the predicted and gold annotations.

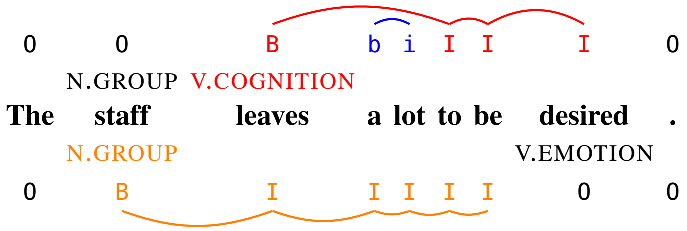


Figure 4: A REVIEWS sentence with MWE and supersense analyses: gold above and hypothetical prediction below. MWE precision of the bottom annotation relative to the top one is $2/5$. (Note that a link between words w_1 and w_2 is "matched" if, in the other annotation, there is a path between w_1 and w_2.) The MWE recall value is $3/4$. Supersense precision and recall are both $1/2$. Combined precision/recall scores add the respective subscores' numerators and denominators: thus, combined precision is $\frac{2+1}{5+2} = 3/7$, and combined recall is $\frac{3+1}{4+2} = 2/3$. Combined F_1 is their harmonic mean, i.e. $12/23$.

and supersense precision ratios by adding their numerators and denominators, and likewise for combined recall (see the example in figure 4).

Within each domain, scores are computed as microaverages. The official tri-domain scores reported here are domain macroaverages: per-domain measures are aggregated with the three domains weighted equally. The main score, tri-domain combined F_1, is the arithmetic mean of the three per-domain combined F_1 scores. (Some system papers report domain microaverages, which give less influence to the TED domain because it is the smallest of the domains in the test set.)

6 Entries and Results

Six teams[18] participated in the task, submitting a total of nine unique system entries prior to the deadline. We give an overview of these systems and analyze their performance.

6.1 Synopsis of approaches

From the **UFRGS&LIF** team (Cordeiro et al., 2016), S106 detects MWEs by heuristic pattern-matching against sequences in the training data, and predicts the most frequent supersense observed for each type in the training data.

From the **UTU** team (Björne and Salakoski, 2016), S211, S254, and S255 match word sequences against a variety of resources and then choose a supersense with an ensemble of classifiers. The method performs reasonably well for supersenses, but is weak at detecting MWEs.

The **UW-CSE** team (Hosseini et al., 2016) experimented with a sequence CRF as well as a double-chained CRF, with separate chains for MWE tags and supersenses, and some parameters shared between them. The closed-condition and open-condition feature sets were drawn from AMALGrAM (Schneider and Smith, 2015). Of the official submissions, S248 used a single-chain CRF and S249 a double-chained CRF. A full comparison demonstrates that the double-chained CRF performs best on the combined measure in both the closed and open conditions.

From the **ICL-HD** team (Kirilin et al., 2016), S214 uses the AMALGrAM sequence tagger (Schneider and Smith, 2015) with an augmented feature set that leverages word embeddings and a knowledge base. The word embedding features, the knowledge base–derived features, and their union all improve over the condition with no new features, with respect to both MWE performance and supersense performance. The best results for the combined measure are obtained with the word embedding features (but not the knowledge base features). The word embeddings are shown to be somewhat complementary to AMALGrAM's Brown cluster features: ablating either reduces performance.

From the **WHUNlp** team (Tang et al., 2016), S108 uses a pipeline where a sequence CRF first identifies

[18] None of the teams included any DiMSUM organizers.

#	System	Team	Score	Resources
1	S214	ICL-HD	57.77	++
	S249	UW-CSE	57.71	++
	S248	UW-CSE	57.10	
2	S106	UFRGS&LIF	50.27	
3	S227	VectorWeavers	49.94	++
4	S255	UTU	47.13	++
5	S211	UTU	46.17	+
	S254	UTU	45.79	
6	S108	WHUNlp	25.71	

Table 4: Main results on the test set. Scores are tri-domain combined F_1 percentages. Resource conditions are described in §6.2.

MWEs, and a maximum entropy classifier then predicts a supersense independently for each lexical expression. Each of these models has a small number of feature templates recording words and POS tags.

From the **VectorWeavers** team (Scherbakov et al., 2016), S227 relies on neural network classifiers to detect MWE boundaries and label supersenses, using features based on word embeddings and syntactic parses. Results show that syntax helps identify MWE boundaries accurately, and that simple incremental composition functions can help construct useful MWE representations.

6.2 Overall results

The main results appear in table 4. The first column of table 4 gives the ranking of the systems. Several systems may share a rank if they do not produce significantly different predictions, as detailed below. The score is the combined supersense and MWE measure, macroaveraged over the three test set domains as described above. The final column indicates the resource condition: systems entered in the open condition (all resources allowed) are designated "++"; "+" indicates the more restricted semi-supervised closed condition, while the remaining systems are in the closed condition (most restrictive). Details of the resource conditions and scoring appear in §5.

Ranking and significance. The overall best scoring system, with a combined measure of 57.77%, is S214. The competition, however, is close: S249 scored 57.71%, and S248 obtained a combined score of 57.10%. To check whether the predictions of the systems are significantly different from each other, we ran McNemar's test, a paired test that operates directly on the predicted system output. A consequence of this is that we do not directly test whether the computed *scores* are significantly different from each other, only whether the *predictions* are.

Submission	Reviews	TED	Tweets
Multiword expressions			
S106	49.57	56.76	51.16
S108	26.39	33.44	34.18
S211 +	9.07	18.28	15.76
S214 ++	53.37	**57.14**	59.49
S227 ++	36.18	41.76	39.32
S248	53.96	52.35	54.48
S249 ++	**54.80**	53.48	**61.09**
S254	7.05	16.30	6.34
S255 ++	8.68	20.11	15.50
Supersenses			
S106	50.93	49.61	49.20
S108	25.82	24.68	24.63
S211 +	52.00	51.40	49.95
S214 ++	**57.66**	**60.06**	55.99
S227 ++	51.36	52.00	51.70
S248	57.19	59.11	56.82
S249 ++	57.00	59.17	**57.46**
S254	52.68	51.44	49.66
S255 ++	51.98	53.28	51.11
Combined score			
S106	50.71	50.57	49.54
S108	25.86	25.39	25.87
S211 +	46.19	47.90	44.42
S214 ++	**56.98**	**59.71**	56.63
S227 ++	49.25	50.82	49.74
S248	56.66	58.26	56.38
S249 ++	56.61	58.33	**58.18**
S254	46.57	47.82	42.99
S255 ++	46.15	49.81	45.44

Table 5: Per-domain evaluation results. Figures are F_1 percentages. The best value in each section and column is in bold. Refer to table 4 for the identities of the systems.

According to McNemar's test, the predictions of the highest-ranking and the next-highest-ranking system are not significantly different at $p < .05$. The third highest ranking system performs significantly worse than the top system, but is *not* significantly different from the second-place system. We therefore decided to rank all three systems together. In general, adjacent entries in the sorted scoring table are ranked together if the difference between them is not statistically significant according to the test.

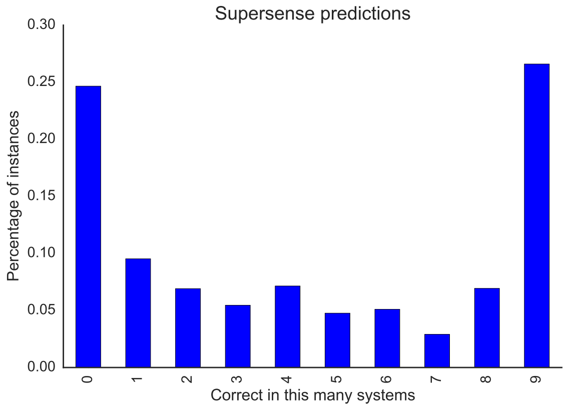


Figure 5: Number of systems predicting the correct supersense (for tokens where there is a gold supersense).

Drilling down. Table 5 offers a more detailed breakdown by domain and subscore (MWEs vs. supersenses vs. combined). The best scores are about 57% for both MWEs and supersenses. Systems S214 and S249 are the clear winners: the former is better in the surprise TED domain—particularly TED MWEs (by nearly 4 points). The latter is slightly better in TWEETS, and the systems are quite close in REVIEWS (the domain with the most training data).

S214 and S249 were in the open condition, taking advantage of additional resources. The best system in the closed condition is S248, which is very similar to S249—and recall that its predictions, overall, are not statistically worse. Table 5 reveals one striking difference, however: in MWE scores for TWEETS, S249 bests S248 by nearly 7 points.

When scores in the 3 domains are compared for each system, there is surprisingly little difference overall. We expected that the TED domain would be most difficult because it is not represented in the training data, but the scores in table 5 give no clear indication that this is the case. Perhaps systems escaped domain bias because the training data included two highly divergent genres; or perhaps other aspects of the data sets (e.g., topic) matter more for this task than differences in genre.

6.3 Easy and hard decisions

Overall, the results clearly show that the joint supersense and MWE tagging task is not yet resolved. Given the wide range of participating systems and previous work, it is reasonable to assume that the task itself is not easy. On the other hand, it is not *uniformly* hard. In fact, some decisions are relatively easy, in the sense that most or all systems get them right; whereas others are hard, in that none or very few systems produce the correct answer. Figure 5 explores this for the supersense-tagging subtask. The tallest bars are near the left and right sides of the graph, representing the hard and easy instances, respectively. Hard instances account for about 25% of instances where the gold data has a supersense, which also puts an upper bound on any system combination. Even an oracle system allowed to choose the best prediction for each instance from among all the systems would still not push the accuracy above 75%.

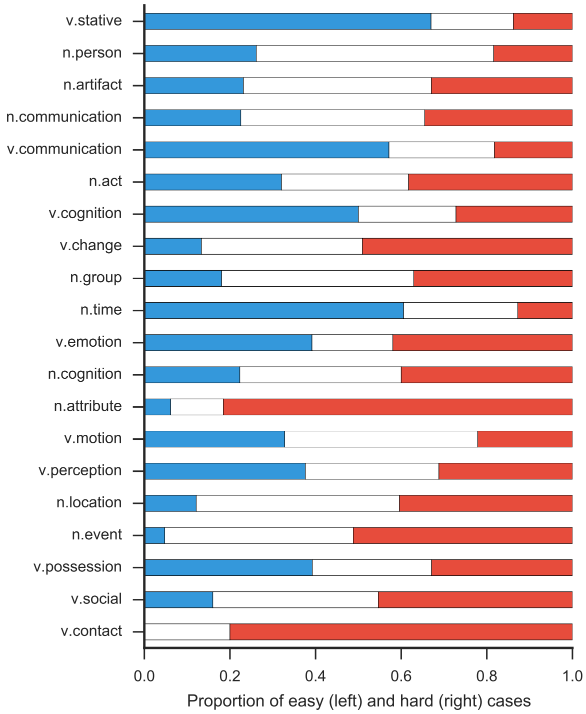


Figure 6: Easy and hard supersense decisions. Shown in blue in the left side of the plot is the proportion of instances of the given supersense type where at most one system gave the *wrong* answer. On the right side in red is the corresponding figure where at most one system gave the *right* answer. Supersenses are sorted by corpus frequency.

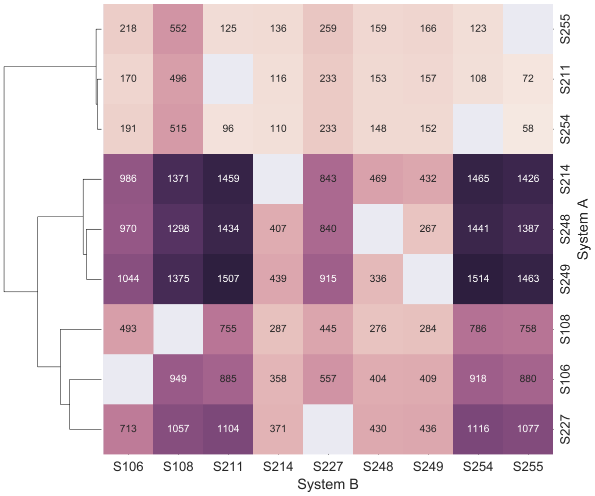
The distribution of easy and hard instances varies a lot between labels, though. As shown for supersenses in figure 6, individual labels range from the fairly easy (e.g. V.STATIVE and V.COMMUNICATION) to the more difficult (e.g. N.ATTRIBUTE and V.CONTACT). The most common supersense, V.STATIVE, is easy because it has few distinct lexical forms (the ten most common lemmas make up more than 77% of the instances). Examples of V.STATIVE lemmas include *be*, *have*, *use*, and *get*.

Supersenses may be difficult for more than one reason. For instance, V.CONTACT—e.g. *deliver*, *receive*, and *take*—has more distinct forms than V.STATIVE and also a more complex mapping between lemmas and supersenses. In contrast, person names, job titles, etc. that should be tagged as N.PERSON are rarely ambiguous with respect to supersense. The main challenge in that case is that the category is open-ended and not in general evident from syntactic structure.

6.4 System correlation

Finally, we examine whether the submitted approaches capture different aspects of the task. I.e., could we produce a better system by combining the individual systems? We cannot estimate this from the results tables, since, combinatorially, there are many ways to obtain a given score. However, we can estimate it from the prediction overlap between systems. The $N \times N$ labeled matrix in figure 7 shows how the N systems relate to each other. Each cell compares the predictions of two systems a and b in the joint supersense and MWE task. The value of a cell $T_{a,b}$ is the number of correct predictions made by a that were not correctly predicted by b. This is an asymmetric measure of predictive similarity. A single low number indicates one out of two things: either the systems are similar, or a is better than b. When the sum $T_{a,b} + T_{b,a}$ is small, the two systems make similar predictions.

Clustering the systems in figure 7 (shown on the left side of the plot) results in groups that correspond to the ranking in table 4. Inside the cluster of systems ranked at 1, the asymmetric predictive advantage ranges between 267 and 469. Lower-ranked systems all have a smaller predictive advantage with respect to the top-ranked systems. The best combination system would thus likely be between two of the rank-1 systems. However, the gains are small, and overall the systems seem to extract the same knowledge, or subsets of the same knowledge, out of the training data.



	S106	S108	S211	S214	S227	S248	S249	S254	S255
S255	218	552	125	136	259	159	166	123	
S211	170	496		116	233	153	157	108	72
S254	191	515	96	110	233	148	152		58
S214	986	1371	1459		843	469	432	1465	1426
S248	970	1298	1434	407	840		267	1441	1387
S249	1044	1375	1507	439	915	336		1514	1463
S108	493		755	287	445	276	284	786	758
S106		949	885	358	557	404	409	918	880
S227	713	1057	1104	371		430	436	1116	1077

Figure 7: System clusters. Each cell compares the predictions of two systems i and j with respect to a gold standard. The value in the i,j-th cell is the number of predictions that i got right but j did not.

7 Conclusion

This task featured a broad-coverage lexical semantic analysis task that combines MWE identification and supersense tagging. The semantic tagset strikes a balance between the extremely difficult fine-grained distinctions in classical WSD, and the restrictiveness of the NER task. To guard against domain bias, we provided training data from two different genres, namely online reviews and tweets, as well as a test-only data set with TED talk transcripts. The training and test data sets are publicly available at `https://github.com/dimsum16/dimsum-data`.

The best scoring systems obtained 57.7% F_1 on a composite measure over the two subtasks of MWE and supersense tagging, averaged over the three test domains. This level of performance suggests that the task is not yet resolved. Furthermore, our error analysis suggests that the submitted systems arrived at similar generalizations from the training data. Substantially improving performance would thus seem to require novel approaches.

Acknowledgments

We are grateful to Sam Gibbon for the lexical semantic annotation, which was generously supported by Carlsberg infrastructure grant No. CF14-0694.

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, *Proc. of LREC*. Valletta, Malta.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA. URL `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T13`.

Jari Björne and Tapio Salakoski. 2016. UTU at SemEval-2016 Task 10: Binary Classification for Expression Detection (BCED). In *Proc. of SemEval*. San Diego, California, USA.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way, editors, *Proc. of EAMT*, pages 261–268. Trento, Italy.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602. Sydney, Australia.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56. Portland, Oregon, USA.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. 2016. UFRGS&LIF at SemEval-2016 Task 10: Rule-based MWE identification and predominant-supersense tagging. In *Proc. of SemEval*. San Diego, California, USA.

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

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2012. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania. URL `http://www.ark.cs.cmu.edu/mheilman/questions/papers/heilman-question-generation-dissertation.pdf`.

Mohammad Javad Hosseini, Noah A. Smith, and Su-In Lee. 2016. UW-CSE at SemEval-2016 Task 10: Detecting multiword expressions and supersenses using double-chained conditional random fields. In *Proc. of SemEval*. San Diego, California, USA.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proc. of WWW*, pages 452–461. Florence, Italy.

Dirk Hovy, Shashank Shrivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proc. of the First Workshop on Metaphor in NLP*, pages 52–57. Atlanta, Georgia, USA.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proc. of ACL-IJCNLP*, pages 483–488. Beijing, China.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What's in a preposition? Dimensions of sense disambiguation for an interesting word class. In *Coling 2010: Posters*, pages 454–462. Beijing, China.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of Twitter. In *Proc. of *SEM*, pages 1–11. Dublin, Ireland.

Angelika Kirilin, Felix Krauss, and Yannick Versley. 2016. ICL-HD at SemEval-2016 Task 10: Improving the detection of minimal semantic units and their meanings with an ontology and word embeddings. In *Proc. of SemEval*. San Diego, California, USA.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proc. of EMNLP*, pages 1001–1012. Doha, Qatar.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Héctor Martínez Alonso, Anders Johannsen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Anna Braasch, Anders Søgaard, and Bolette Sandford Pedersen. 2015. Supersense tagging for Danish. In Beáta Megyesi, editor, *Proc. of NODALIDA*, pages 21–29. Vilnius, Lithuania.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308. Plainsboro, NJ, USA.

Graham Neubig, Katsuhito Sudoh, Yusuke Oda, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2014. The naist-ntt ted talk treebank. In *International Workshop on Spoken Language Translation*.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2015. Universal Dependencies 1.2. URL `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1548`, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of NAACL-HLT*, pages 380–390. Atlanta, Georgia, USA.

Gerhard Paaß and Frank Reichartz. 2009. Exploiting semantic constraints for estimating supersenses with CRFs. In *Proc. of the Ninth SIAM International Conference on Data Mining*, pages 485–496. Sparks, Nevada.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv:1104.2086 [cs]*. URL `http://arxiv.org/abs/1104.2086`.

Davide Picca, Alfio Massimiliano Gliozzo, and Simone Campora. 2009. Bridging languages by SuperSense entity tagging. In *Proc. of NEWS*, pages 136–142. Suntec, Singapore.

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, *Proc. of LREC*, pages 2386–2390. Marrakech, Morocco.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proc. of EACL*, pages 742–751. Gothenburg, Sweden.

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 (CICLing'11)*, volume 6608 of *Lecture Notes in Computer Science*, pages 15–28. Springer, Berlin.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proc. of ACL 2012 Student Research Workshop*, pages 1–6. Jeju Island, Korea.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In

Proc. of the Third ACL Workshop on Very Large Corpora, pages 82–94. Cambridge, MA.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proc. of EMNLP*, pages 1524–1534. Edinburgh, Scotland, UK.

Stefano Dei Rossi, Giulia Di Pietro, and Maria Simi. 2013. Description and results of the SuperSense tagging task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, number 7689 in Lecture Notes in Computer Science, pages 166–175. Springer Berlin Heidelberg.

Andreas Scherbakov, Ekaterina Vylomova, Fei Liu, and Timothy Baldwin. 2016. VectorWeavers at SemEval-2016 Task 10: From incremental meaning to semantic unit (phrase by phrase). In *Proc. of SemEval*. San Diego, California, USA.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Oflazer, and Noah A. Smith. 2013. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. of NAACL-HLT*, pages 661–667. Atlanta, Georgia, USA.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse lexical semantic annotation with supersenses: an Arabic case study. In *Proc. of ACL*, pages 253–258. Jeju Island, Korea.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 455–461. Reykjavík, Iceland.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proc. of NAACL-HLT*, pages 1537–1547. Denver, Colorado.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proc. of The 9th Linguistic Annotation Workshop*, pages 112–123. Denver, Colorado, USA.

Xin Tang, Fei Li, and Donghong Ji. 2016. WHUNlp at SemEval-2016 Task 10: A pilot study in detecting minimal semantic units and their meanings using supervised models. In *Proc. of SemEval*. San Diego, California, USA.

Stephen Tratz and Eduard Hovy. 2010. ISI: Automatic classification of relations between nominals using a maximum entropy classifier. In *Proc. of SemEval*, pages 222–225. Uppsala, Sweden.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proc. of the First Workshop on Metaphor in NLP*, pages 45–51. Atlanta, Georgia, USA.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting English adjective senses with supersenses. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 4359–4365. Reykjavík, Iceland.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. of MUC-6*, pages 45–52. Columbia, Maryland, USA.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. Detecting noun compounds and light verb constructions: a contrastive study. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 116–121. Portland, Oregon, USA.

Veronika Vincze, János Zsibrita, and István Nagy T. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proc. of the Sixth International Joint Conference on Natural Language Processing*, pages 207–215. Nagoya, Japan.

Patrick Ye and Timothy Baldwin. 2007. MELB-YB: Preposition sense disambiguation using rich semantic features. In *Proc. of SemEval*, pages 241–244. Prague, Czech Republic.