Supersense Embeddings: A Unified Model for Supersense Interpretation, Prediction, and Utilization

Lucie Flekova[†] **and Iryna Gurevych**[†‡]
† Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
‡ Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract

Coarse-grained semantic categories such as supersenses have proven useful for a range of downstream tasks such as question answering or machine translation. To date, no effort has been put into integrating the supersenses into distributional word representations. We present a novel joint embedding model of words and supersenses, providing insights into the relationship between words and supersenses in the same vector space. Using these embeddings in a deep neural network model, we demonstrate that the supersense enrichment leads to a significant improvement in a range of downstream classification tasks.

1 Introduction

The effort of understanding the meaning of words is central to the NLP community. The word sense disambiguation (WSD) task has therefore received a substantial amount of attention (see Navigli (2009) or Pal and Saha (2015) for an overview). Words in training and evaluation data are usually annotated with senses taken from a particular lexical semantic resource, most commonly WordNet (Miller, 1995). However, WordNet has been criticized to provide too fine-grained distinctions for end level applications. e.g. in machine translation or information retrieval (Izquierdo et al., 2009). Although some researchers report an improvement in sentiment prediction using WSD (Rentoumi et al., 2009; Akkaya et al., 2011; Sumanth and Inkpen, 2015), the publication bias toward positive results (Plank et al., 2014) impedes the comparison to experiments with the opposite conclusion, and the contribution of WSD to downstream document classification tasks remains "mostly speculative"(Ciaramita and Altun, 2006), which can be attributed to the too subtle sense distinctions (Navigli, 2009). This is why *supersenses*, the coarse-grained word labels based on WordNet's (Fellbaum, 1998) lexicographer files, have recently gained attention for text classification tasks. Supersenses contain 26 labels for nouns, such as ANIMAL, PERSON or FEELING and 15 labels for verbs, such as COMMUNICATION, MOTION or COGNITION. Usage of supersense labels has been shown to improve dependency parsing (Agirre et al., 2011), named entity recognition (Marrero et al., 2009; Rüd et al., 2011), non-factoid question answering (Surdeanu et al., 2011), question generation (Heilman, 2011), semantic role labeling (Laparra and Rigau, 2013), personality profiling (Flekova and Gurevych, 2015), semantic similarity (Severyn et al., 2013) and metaphor detection (Tsvetkov et al., 2013).

An alternative path to semantic interpretation follows the distributional hypothesis (Harris, 1954). Recently, word vector representations learned with neural-network based language models have contributed to state-of-the-art results on various linguistic tasks (Bordes et al., 2011; Mikolov et al., 2013b; Pennington et al., 2014; Levy et al., 2015).

In this work, we present a novel approach for incorporating the supersense information into the word embedding space and propose a new methodology for utilizing these to label the text with supersenses and to exploit these joint word and supersense embeddings in a range of applied text classification tasks. Our contributions in this work include the following:

- We are the first to provide a joint word- and supersense-embedding model, which we make publicly available[1] for the research community. This provides an insight into the word and supersense positions in the vector space

[1] https://github.com/UKPLab/acl2016-supersense-embeddings

through similarity queries and visualizations, and can be readily used in any word embedding application.

- Using this information, we propose a supersense tagging model which achieves competitive performance on recently published social media datasets.
- We demonstrate how these predicted supersenses and their embeddings can be used in a range of text classification tasks. Using a deep neural network architecture, we achieve an improvement of 2-6% in accuracy for the tasks of sentiment polarity classification, subjectivity classification and metaphor prediction.

2 Related Work

2.1 Semantically Enhanced Word Embeddings

An idea of combining the distributional information with the expert knowledge is attractive and has been newly pursued in multiple directions. One of them is creating the word sense or synset embeddings (Iacobacci et al., 2015; Chen et al., 2014; Rothe and Schütze, 2015; Bovi et al., 2015). While the authors demonstrate the utility of these embeddings in tasks such as WSD, knowledge base unification or semantic similarity, the contribution of such vectors to downstream document classification problems can be challenging, given the fine granularity of the WordNet senses (cf. the discussion in Navigli (2009)). As discussed above, supersenses have been shown to be better suited for carrying the relevant amount of semantic information. An alternative approach focuses on altering the objective of the learning mechanism to capture relational and similarity information from knowledge bases (Bordes et al., 2011; Bordes et al., 2012; Yu and Dredze, 2014; Bian et al., 2014; Faruqui and Dyer, 2014; Goikoetxea et al., 2015). While, in principle, supersenses could be seen as a relation between a word and its hypernym, to our knowledge they have not been explicitly employed in these works. Moreover, an important advantage of our explicit supersense embeddings compared to the retrained vectors is their direct interpretability.

2.2 Supersense Tagging

Supersenses, also known as lexicographer files or semantic fields, were originally used to organize lexical-semantic resources (Fellbaum, 1990). The supersense tagging task was introduced by Ciaramita and Johnson (2003) for nouns and later expanded for verbs (Ciaramita and Altun, 2006). Their state-of-the-art system is trained and evaluated on the SemCor data (Miller et al., 1994) with an F-score of 77.18%, using a hidden Markov model. Since then, the system, resp. its reimplementation by Heilman[2], was widely used in applied tasks (Agirre et al., 2011; Surdeanu et al., 2011; Laparra and Rigau, 2013). Supersense taggers have then been built also for Italian (Picca et al., 2008), Chinese (Qiu et al., 2011) and Arabic (Schneider et al., 2013). Tsvetkov et al. (2015) proposes the usage of SemCor supersense frequencies as a way to evaluate word embedding models, showing that a good alignment of embedding dimensions to supersenses correlates with performance of the vectors in word similarity and text classification tasks. Recently, Johannsen et al. (2014) introduced a task of multiword supersense tagging on Twitter. On their newly constructed dataset, they show poor domain adaptation performance of previous systems, achieving a maximum performance with a search-based structured prediction model (Daumé III et al., 2009) trained on both Twitter and SemCor data. In parallel, Schneider and Smith (2015) expanded a multiword expression (MWE) annotated corpus of online reviews with supersense information, following an alternative annotation scheme focused on MWE. Similarly to Johannsen et al. (2014), they find that SemCor may not be a sufficient resource for supersense tagging adaption to different domains. Therefore, in our work, we explore the potential of using an automatically annotated Babelfied Wikipedia corpus (Scozzafava et al., 2015) for this task.

3 Building Supersense Embeddings

To learn our embeddings, we adapt the freely available sample of 500k articles of Babelfied English Wikipedia (Scozzafava et al., 2015). To our knowledge, this is one of the largest published and evaluated sense-annotated corpora, containing over 500 million words, of which over 100 million are annotated with Babel synsets, with an estimated synset annotation accuracy of 77.8%. Few other automatically sense-annotated Wikipedia corpora are available (Jordi Atserias and Attardi, 2008; Reese et

[2] https://github.com/kutschkem/SmithHeilmann_fork/tree/master/MIRATagger

.1	*About 10.9% of families were below the poverty line, including 13.6% of those under age 18.*
2	*About 10.9% of* N.GROUP *were below the* N.POSSESSION V.CHANGE *13.6% of those under* N.ATTRIBUTE *18.*
3	*About 10.9% of* FAMILIES_N.GROUP *were below the* POVERTY_LINE_N.POSSESSION INCLUDING_V.CHANGE *13.6% of those under* AGE_N.ATTRIBUTE *18.*

Table 1: Example of plain (1), generalized (2) and disambiguated (3) Wikipedia

al., 2010). However, their annotation quality was assessed only on the training domain and as Atserias et al. state (p.2316): *"Wikipedia text differs significantly ... from the corpora used to train the taggers ... Therefore the quality of these NLP processors is considerably lower than the results of the evaluation in-domain."*

We map the Babel synsets to WordNet 3.0 synsets (Miller, 1995) using the BabelNet API (Navigli and Ponzetto, 2012), and map these synsets to their corresponding WordNet's supersense categories (Miller, 1990; Fellbaum, 1990). For the nested named entities, only the largest BabelNet span is considered, hence there are no nested supersense labels in our data. In this manner we obtain an alternative Wikipedia corpus, where each word is replaced by its corresponding supersense (see Table 1, second row) and another alternative corpus where each word has its supersense appended (Table 1, third row). Using the Gensim (Řehůřek and Sojka, 2010) implementation of Word2vec (Mikolov et al., 2013a), we applied the skip-gram model with negative sampling on these three Wikipedia corpora jointly (i.e., on the rows 1, 2 and 3 in Table 1) to produce continuous representations of words, supersense-disambiguated words and standalone supersenses in one vector space based on the distributional information obtained from the data. [3] The benefits of learning this information jointly are threefold:

1. Vectorial representations of the original words are altered (compared to training on text only), taking into account the similarity to supersenses in the vector space

2. Standalone supersenses are positioned in the vector space, enabling insightful similarity queries between words and supersenses, esp. for unannotated words

3. Disambiguated word+supersense vectors of annotated words can be employed similarly to sense embeddings (Iacobacci et al., 2015; Chen et al., 2014) to improve downstream tasks and serve as input for supersense disambiguation or contextual similarity systems

In the following, the designation WORDS denotes the experiments with the word embeddings learned on plain Wikipedia text (as in row 1 of Table 1) while the designation SUPER denotes the experiments with the word embeddings learned jointly on the supersense-enriched Wikipedia (i.e., rows 1, 2 and 3 in Table 1 together).

4 Qualitative Analysis

4.1 Verb Supersenses

Table 2 shows the most similar word vectors to each of the verb supersense vectors using cosine similarity. Note that while no explicit part-of-speech information is specified, the most similar words hold both the semantic and syntactic information - most of the assigned words are verbs.

VERBS	
BODY	wearing, injured, worn, wear, wounded, bitten, soaked, healed, cuffed, dressed
CHANGE	changed, started, added, dramatically, expanded drastically, begun, altered, shifted, transformed
COGNITION	known, thought, consider, regarded, remembered attributed, considers, accepted, believed, read
COMMUNICATION	stated, said, argued, jokingly, called, noted, suggested, described, claimed, referred
COMPETITION	won, played, lost, beat, scored defeated, win, competed, winning, playing
CONSUMPTION	feed, fed, employed, based, hosted feeds, utilized, applied, provided, consumed
CONTACT	thrown, set, carried, opened, laid pulled, placed, cut, dragged, broken
CREATION	produced, written, created, designed, developed directed, built, published, penned, constructed
EMOTION	want, felt, loved, wanted, delighted disappointed, feel, like, saddened, thrilled
MOTION	brought, led, headed, returned, followed left, turned, sent, travelled, entered
PERCEPTION	seen, shown, revealed, appeared, appears shows, noticed, see, showing, presented
POSSESSION	received, obtained, awarded, acquired, provided donated, gained, bought, found, sold
SOCIAL	appointed, established, elected, joined, assisted led, succeeded, encouraged, initiated, organized
STATIVE	included, held, includes, featured, served, represented, referred, holds, continued, related
WEATHER	glow, emitted, ignited, flare, emitting smoke, fumes, sunlight, lit, darkened

Table 2: Top 10 most similar word embeddings for verb supersense vectors

[3] The embeddings are learned using skip-gram as training algorithm with downsampling of 0.001 higher-frequency words, negative sampling of 5 noise words, minimal word frequency of 100, window of size 2 and alpha of 0.025, using 10 epochs to produce 300-dimensional vectors. Our experiments with less dimensions and with the CBOW model performed worse.

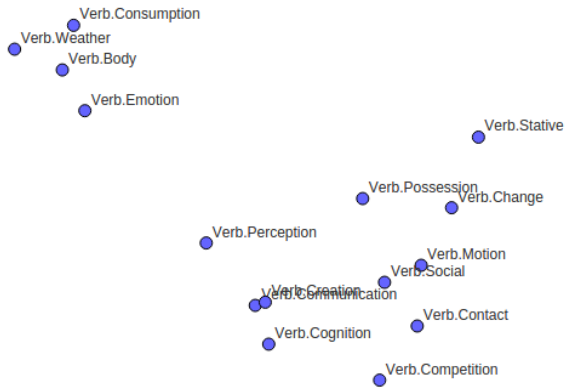

Figure 1: Verb supersense embeddings visualized in the vector space (t-SNE)

Furthermore, using a large corpus such as Wikipedia conveniently reduces the current need of lemmatization for supersense tagging, as the words are sufficiently represented in all their forms. The most frequent error originates from assigning the adverbs to their related verb categories, e.g. *jokingly* to COMMUNICATION and *drastically* to CHANGE. Such information, however, can be beneficial for context analysis in supersense tagging.

Figure 1 displays the verb supersenses using the t-distributed Stochastic Neighbor Embedding (Van der Maaten and Hinton, 2008), a technique designed to visualize structures in high-dimensional data. While many of the distances are probable to be dataset-agnostic, such as the proximity of BODY, CONSUMPTION and EMOTION, other appear emphasized by the nature of Wikipedia corpus, e.g. the proximity of supersenses COMMUNICATION and CREATION or SOCIAL and MOTION, as can be explained by table 2 (see e.g. *led* and *followed*).

Figure 2: Noun supersense embeddings (t-SNE)

4.2 Noun Supersenses

Table 3 displays the most similar word embeddings for noun supersenses. In accordance with previous work on suppersense tagging (Ciaramita and Altun, 2006; Schneider et al., 2012; Johannsen et al., 2014), the assignments of more specific supersenses such as FOOD, PLANT, TIME or PERSON are in general more plausible than those for abstract concepts such as ACT, ARTIFACT or COGNITION. The same is visible in Figure 2, where these supersense embeddings are more central, with closer neighbors. In contrast to the observations by Schneider et al. (2012) and Johannsen et al. (2014), the COMMUNICATION supersense appears well defined, likely due to the character of Wikipedia.

NOUNS	
ACT	participation, activities, involvement, undertaken ongoing, conduct, efforts, large-scale, success
ANIMAL	peccaries, capybaras, frogs, echidnas, birds marmosets, rabits, hatchling, ciconiidae, species
ARTIFACT	wooden, two-floor, purpose-built, installed, wall fittings, turntable, racks, wrought-iron, ceramic, stone
ATTRIBUTE	height, strength, age, versatility, hardness power, fluidity, mastery, brilliance, inherent
BODY	abdomen, bone, femur, anterior, forearm femoral, skin, neck, muscles, thigh
COGNITION	ideas, concepts, empirical, philosophy, knowledge, epistemology, analysis, atomistic, principles
COMMUNICATION	written, excerpts, text, music, excerpted, translation, lyrics, subtitle, transcription, words
EVENT	sudden, death, occurred, event, catastrophic unexpected, accident, victory, final, race
FEELING	sadness, love, sorrow, frustration, disgust anger, affection, feelings, grief, fear
FOOD	cheese, butter, coffee, milk, yogurt dessert, meat, bread, vegetables, sauce
GROUP	members, school, phtheochroa, ypsolophidae pitcairnia, cryptanthus, group, division, schools
LOCATION	northern, southern, northeastern, area, south capital, town, west, region, city
MOTIVE	motivation, reasons, rationale, justification, motive justifications, motives, incentive, desire, why
OBJECT	river, valley, lake, hills, floodplain lakes, rivers, mountain, estuary, ocean
PERSON	greatgrandfather, son, nephew, son-in-law, father halfbrother, brother, who, mentor, fellow
PHENOMENON	wind, forces, self-focusing, radiation, ionizing result, intensity, gravitational, dissipation, energy
PLANT	fruit, fruits, magnifera, sativum, flowers caesalpinia, shrubs, trifoliate, vines, berries
POSSESSION	property, payment, money, payments, taxes tax, cash, fund, pay, $100
PROCESS	growth, decomposition, oxidative, mechanism rapid, reaction, hydrolysis, inhibition, development
QUANTITY	miles, square, meters, kilometer, cubic, ton, number, megabits, volume, kilowatthours
RELATION	southeast, southwest, northeast, northwest, east portion, link, correlation, south, west
SHAPE	semicircles, right-angled, concave, parabola, ellipse, angle, circumcircle, semicircle, lines
STATE	chronic, condition, debilitating, problems, health worsening, illness, illnesses, exacerbation, disease
SUBSTANCE	magnesium, zinc, silica, manganese, sulfur oxide, sulphate, phosphate, salts, phosphorus
TIME	september, december, november, july, april january, august, february, year, days
TOPS	time, group, event, person, groups individuals, events, animals, individual, plant

Table 3: Top 10 most similar word embeddings for noun supersense vectors

4.3 Word Analogy and Word Similarity Tasks

We also assess the changes between the individual word embeddings learned on plain Wikipedia text (WORDS) and jointly with the supersense-enriched Wikipedia (SUPER). With this aim we perform two standard embedding evaluation tasks: word similarity and word analogy.

Mikolov et al. (2013b) introduce a word analogy dataset containing 19544 analogy questions that can be answered with word vector operations (*Paris is to France as Athens are to...?*). The questions are grouped into 13 categories. Table 4 presents our results. Word vectors trained in the SUPER setup achieve better results on groups related to entities, e.g. Family Relations and Citizen to State questions, where the PERSON and LOCATION supersenses can provide additional information to reduce noise. At the same time, performance on questions such as Opposites or Plurals drops, as this information is pushed to the background. Enriching our data with the recently proposed adjective supersenses (Tsvetkov et al., 2014) could be of interest for these categories.

Group/Vectors:	WORDS	SUPER
Capitals - common	91.1	94.7±0.99
Capitals - world	87.6	89.5±0.69
City in state	65.2	65.7±1.03
Nationality to state	94.5	95.2±0.58
Family relations	93.0	94.4±1.28
Opposites	56.7	54.6±3.21
Plurals	89.4	86.4±1.08
Comparatives	90.6	90.4±0.85
Superlatives	79.4	79.6±1.83
Adjective to adverb	20.2	22.2±1.53
Present to participle	64.2	64.6±1.57
Present to past	60.0	59.2±1.30
3rd person verbs	84.3	82.1±1.44
Total	75.0	76.0±0.28

Table 4: Accuracy and standard error on analogy tasks. Tasks related to noun supersense distinctions show the tendency to improve, while syntax-related information is pushed to the background. In most cases, however, the difference is not significant.

Without explicitly exploiting the sense infromation, we compare the performance of our text-trained (WORDS) to our jointly trained (SUPER) word vectors on the following word similarity datasets: WordSim353-Similarity (353-S) and WordSim353-Relatedness (353-R) (Agirre et al., 2009), MEN dataset (Bruni et al., 2014), RG-65 dataset (Rubenstein and Goodenough, 1965) and MC-30 (Miller and Charles, 1991).

Data:	MEN	353-S	353-R	RG-65	MC-30
WORDS	73.18	76.93	62.11	79.13	79.49
SUPER	74.26	78.63	61.22	79.75	80.94

Table 5: Performance of our vectors (Spearman's ρ) on five similarity datasets. Results indicate a trend of better performance of vectors trained jointly with supersenses.

The word embeddings for words trained jointly with supersenses achieve higher performance than those trained solely on the same text without supersenses on 4 out of 5 tasks (Table 5). In addition, the explicit supersense information could be further exploited, similarly to previous sense embedding works (Iacobacci et al., 2015; Rothe and Schütze, 2015; Chen et al., 2014). Furthermore, note that while we report the performance of our embeddings on the word similarity tasks for completeness, there has been a substantial discussion on seeking alternative ways to quantify embedding quality with the focus on their purpose in downstream applications (Li and Jurafsky, 2015; Faruqui et al., 2016). Therefore, in the remainder of this paper we explore the usefulness of supersense embeddings in text classification tasks.

5 Building a Supersense Tagger

The task of predicting supersenses has recently regained its popularity (Johannsen et al., 2014; Schneider and Smith, 2015), since supersenses provide disambiguating information, useful for numerous downstream NLP tasks, without the need of tedious fine-grained WSD. Exploiting our joint embeddings, we build a deep neural network model to predict supersenses on the Twitter supersense corpus created by Johannsen et al. (2014), based on the Twitter NER task (Ritter et al., 2011), using the same training data as the authors. [4][5] The datasets follow the token-level annotation which combines the B-I-O flags (Ramshaw and Marcus, 1995) with the supersense class labels to represent the multiword expression segmentation and supersense labeling in a sentence.

5.1 Experimental Setup

We implement a window-based approach with a multi-channel multi-layer perceptron model using

[4] https://github.com/kutschkem/SmithHeilmann_fork/tree/master/MIRATagger/data

[5] https://github.com/coastalcph/supersense-data-twitter

the Theano framework (Bastien et al., 2012). With a sliding window of size 5 for the sequence learning setup we extract for each word the following seven feature vectors:

1. 300-dimensional word embedding,
2. 41 cosine similarities of the word to each standalone supersense embedding,
3. 41 cosine similarities of the word to each of its *word*_SUPERSENSE embeddings,
4. fixed vector of frequencies of each supersense in Wikipedia, in order to simulate the MFS backoff strategy,
5. for the given word, the frequency of each *word*_SUPERSENSE in our Wikipedia corpus,
6. part-of-speech information as a unit vector,
7. casing information as a 3-dimensional (upper/lower/mixed) unit vector

After a dropout regularization, the embedding sets are flattened, concatenated and fed into fully connected dense layers with a rectified linear unit (ReLU) activation function and a final softmax.

5.2 Supersense Prediction

We evaluate our system on the same Twitter dataset with provided training and development (`Tw-R-dev`) set and two test sets: `Tw-R-eval`, reported by Johannsen et al. as *RITTER*, and `Tw-J-eval`, reported by Johannsen et al. as *INHOUSE*. Our results are shown in table 6 and compared to results reported in previous work by Johannsen et al. (2014), with two additional baselines: The SemCor system of Ciaramita and Altun (2006) and the most frequent sense. Our system achieves comparable performance to the best previously used supervised systems, without using any explicit gazetteers.

To get an intuition.[6] of how the individual feature vectors contribute to the prediction, we perform an ablation test by removing one feature group at a time. The biggest performance drop in the F-score (2.7–5.4) occurs when removing the the part of speech information, followed by the supersense similarity features and supersense frequency priors (0.2–3.0). The casing information has only a minor contribution to Twitter supersense tagging (0–0.9).

System/Data:	Tw-R-dev	Tw-R-eval	Tw-J-eval
Baseline and upper bound			
Most frequent sense	47.54	44.98	38.65
Inter-annotator agreement		69.15	61.15
SemCor-trained systems			
(Ciaramita and Altun, 2006)†	48.96	45.03	39.65
Searn (Johannsen et al., 2014)	56.59	50.89	40.50
HMM (Johannsen et al., 2014)	57.14	50.98	41.84
Ours Semcor	54.47	50.30	35.61
Twitter-trained systems			
Searn (Johannsen et al., 2014)	67.72	57.14	42.42
HMM (Johannsen et al., 2014)	60.66	51.40	41.60
Ours Twitter (all features)	61.12	57.16	41.97
Ours Twitter no casing	61.06	56.20	41.13
Ours Twitter no similarities	63.47	56.78	39.44
Ours Twitter no frequencies	61.10	57.32	39.02
Ours Twitter no part-of-speech	57.08	54.45	36.50
Ours Twitter no word embed.	57.57	53.43	34.91

Table 6: Weighted F-score performance on supersense prediction for the development set and two test sets provided by Johannsen et al. (2004). Our system performs comparably to state-of-the-art systems.
† For the system of Ciaramita et al, the publicly avaliable reimplementation of Heilman was used

6 Using Supersense Embeddings in Document Classification Tasks

Word sense disambiguation is to some extent an artificial stand-alone task. Despite its popularity, its contribution to downstream document classification tasks remains rather limited, which might be attributed to the complexity of document preprocessing and the errors cumulated along the pipeline. In this section, we demonstrate an alternative, deep learning approach, in which we process the original text in parallel to the supersense information. The model can then flexibly learn the usefulness of provided input. We demonstrate that the model extended with supersense embeddings outperforms the same model using only word-based features on a range of classification tasks.

6.1 Experimental Setup

Both Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) are state-of-the-art semantic composition models for a variety of text classification tasks (Kim, 2014; Li et al., 2015; Johnson and Zhang, 2014). Recently, their combinations have been proposed, achieving an unprecedented performance (Sainath et al., 2015). We extend the CNN-LSTM approach from the publicly available

[6] Intuition, since there are many additional aspects that may affect the performance. For example, we keep the network parameters fixed for the ablation, although the feature vectors are of different lengths. Furthermore, our model performs a concatenation of the feature vectors, hence only the ablation extended to all possible permutations would verify the feature order effect.

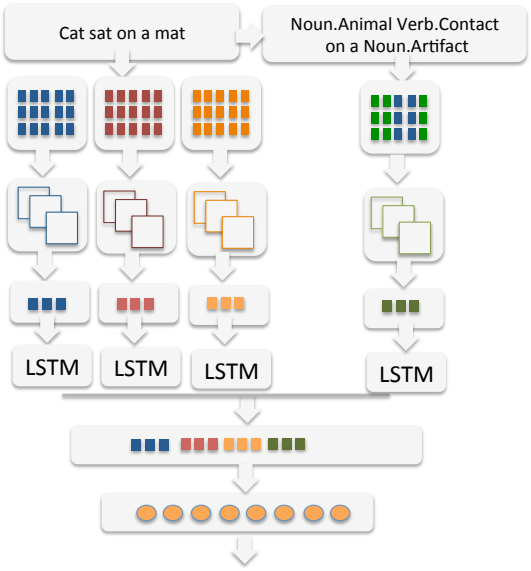

Figure 3: Network architecture. Each of the four different embedding channels serves as input to its CNN layer, followed by an LSTM layer. Afterwards, the outputs are concatenated and fed into a dense layer.

Keras demo[7], into which we incorporate the supersense information. Figure 3 displays our network architecture. First, we use three channels of word embeddings on the plain textual input. The first channel are the 300-dimensional word embeddings obtained from our enriched Wikipedia corpus. The second embedding channel consists of 41-dimensional vectors capturing the cosine similarity of the word to each supersense embedding. The third channel contains the vector of relative frequencies of the word occurring in the enriched Wikipedia together with its supersense, i.e. providing the background supersense distribution for the word. Each of the document embeddings is then convoluted with the filter size of 3, followed by a pooling layer of length 2 and fed into a long-short-term-memory (LSTM) layer. In parallel, we feed as input a processed document text, where the words are replaced by their predicted supersenses. Given that we have the Wikipedia-based supersense embeddings in the same vector space as the word embeddings, we can now proceed to creating the 300-dimensional embedding channel also for the supersense text. As in the plain text channels, we feed also these embeddings into the convolutional and LSTM layers in a similar fashion. Afterwards, we concatenate all LSTM outputs and feed them into a standard fully connected neural network layer, followed by the sigmoid for the binary output. The following subsections discuss our results on a range of classification tasks: subjectivity prediction, sentiment polarity classification and metaphor detection.

6.2 Sentiment Polarity Classification

Sentiment classification has been a widely explored task which received a lot of attention. The Movie Review dataset, published by Pang and Lee (2005)[8], has become a standard machine learning benchmark task for binary sentence classification. Socher et al. (2011) address this task with recursive autoencoders and Wikipedia word embeddings, later improving their score using recursive neural network with parse trees (Socher et al., 2012). Competitive results were achieved also by a sentiment-analysis-specific parser (Dong et al., 2015), with a fast dropout logistic regression (Wang and Manning, 2013), and with convolutional neural networks (Kim, 2014). Table 7 compares these approaches to our results for a 10-fold crossvalidation with 10% of the data withheld for parameter tuning. The line *WORDS* displays the performance using only the leftmost part of our architecture, i.e. only the text input with our word embeddings. The line *SUPER* shows the result of using the full supersense architecture. As it can be seen from the table, the supersense features improve the accuracy by about 2%. Both systems are significantly different ($p < 0.01$), using the McNemar's test.

System	Accuracy
Socher et al. (2011)	77.7
Socher et al. (2012)	79.0
Wang and Manning (2013)	79.1
Dong et al. (2015)	79.5
Kim (2014)	81.5
WORDS	79.4
SUPER	**81.7**±0.37

Table 7: 10-fold cross-validation accuracy and standard error of our system and as reported in previous work for the sentiment classification task on Pang and Lee (2005) movie review data

A detailed analysis of the supersense-tagged data and the classification output revealed that supersenses help to generalize over rare terms. Noun

[7] https://github.com/fchollet/keras/blob/master/examples/imdb_cnn_lstm.py

[8] http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html

Positive reviews	
Text	Supersenses
beating the austin powers film at their own game , this blaxploitation spoof downplays the raunch in favor of gags that rely on the strength of their own cleverness as oppose to the extent of their outrageousness .	verbstative the nounlocation nouncognition nounartifact at their own nouncommunication , this nounact nouncommunication verbstative the nouncognition in nouncommunication of that verbcognition on the nouncognition of their own nouncognition as verbcommunication to the nounevent of their nounattribute .
there is problem with this film that even 3 oscar winner ca n't overcome , but it 's a nice girl-buddy movie once it get rock-n-rolling .	there verbstative nouncognition with this nouncommunication that even 3 nounevent nounperson ca n't verbemotion , but it verbstative a nice girl-buddy nouncommunication once it verbstative rock-n-rolling
godard 's ode to tackle life 's wonderment is a rambling and incoherent manifesto about the vagueness of topical excess . in praise of love remain a ponderous and pretentious endeavor that 's unfocused and tediously exasperating .	nounperson nouncommunication to verbstative nouncognition 's nouncognition verbstative a rambling and incoherent nouncommunication about the nounattribute of topical excess . in nouncognition of nouncognition verbstative a ponderous and pretentious nounact that verbstative unfocused and tediously exasperating
Negative reviews	
Text	Supersenses
the action scene has all the suspense of a 20-car pileup , while the plot hole is big enough for a train car to drive through – if kaos have n't blow them all up .	the nounact nounlocation verbstative all the nouncognition of a 20-car nouncognition , while the nounlocation verbstative big enough for a nounartifact nounartifact to verbmotion through – if nounperson have n't verbcommunication them all up .
the scriptwriter is no less a menace to society than the film 's character .	the nounperson verbstative no less nounstate to noungroup than the nouncommunication nounperson .
a very slow , uneventful ride around a pretty tattered old carousel .	a very slow , uneventful nounact around a pretty tattered old nounartifact .
the milieu is wholly unconvincing . . . and the histrionics reach a truly annoying pitch .	the nouncognition verbstative wholly unconvincing and the nouncommunication verbstative a truly annoying nounattribute .

Table 8: Example of documents classified incorrectly with word embeddings and correctly with word and supersense embeddings on Pang and Lee (2005) movie review data.

concepts such as GROUP, LOCATION, TIME and PERSON appear somewhat more frequently in positive reviews while certain verb supersenses such as PERCEPTION, SOCIAL and COMMUNICATION are more frequent in the negative ones. On the other hand, the supersense tagging introduces additional errors too - for example the director's *cut* is persistently classified into FOOD.

Table 8 shows an example of positive and negative reviews which were consistently (5x in repeated experiments with different random seeds) classified incorrectly with word embeddings and classified correctly with supersense embeddings. Often the wit of unusual expressions is lost for the benefit of generalization. Some improvements appear to be a result of replacing proper names by NOUN.PERSON.

6.3 Subjectivity Classification

Pang and Lee (2004) demonstrate that the subjectivity detection can be a useful input for a sentiment classifier. They compose a publicly available dataset[9] of 5000 subjective and 5000 objective sentences, classifying them with a reported accuracy of 90-92% and further show that predicting this information improves the end-level sentiment classification on a movie review dataset. Kim (2014) and Wang and Manning (2013) further improve the performance through different machine learning methods. Supersenses are a natural candidate for subjectivity prediction, as we hypothesize that the nouns and verbs in the subjective and objective sentences often come from different semantic classes (e.g. VERB.FEELING vs. VERB.COGNITION). We employ the same architecture as in previous task, automatically annotating the words in the documents with their supersenses. Our results are reported in Table 9. The supersenses (SUPER) provide an additional information, improving the model performance by up to 2% over word embeddings (WORDS). The difference between both systems is significant. Based on a manual error analysis, the supersense information contributes here in a similar manner as in the previous case. Subjective sentences contain more verbs of supersense PERCEPTION, while objective ones more frequently feature the supersenses POSSESSION and SOCIAL. Nouns in the subjective category are characterized by supersenses COMMUNICATION and ATTRIBUTE, while in objective ones the PERSON and POSSESSION are more frequent.

System	Accuracy
SVM (Pang and Lee, 2004)	90.0
NB (Pang and Lee, 2004)	92.0
CNN (Kim, 2014)	93.4
F-Dropout (Wang and Manning, 2013)	93.6
MV-CNN (Zhang et al., 2016)	93.9
WORDS	92.1
SUPER	**93.9**$\pm$0.26

Table 9: 10-fold cross-validation accuracy and standard error of our system and as reported in previous work for binary classification on the subjectivity dataset of Pang and Lee (2004)

[9] https://www.cs.cornell.edu/people/pabo/movie-review-data/

6.4 Metaphor Identification

Supersenses have recently been shown to provide improvements in metaphor prediction tasks (Gershman et al., 2014), as they hold the information of coarse semantic concepts. Turney et al. (2011) explore the task of discriminating literal and metaphoric adjective-noun expressions. They report an accuracy of 79% on a small dataset rated by five annotators. Tsvetkov et al. (2013) pursue this work further by constructing and publishing a dataset of 985 literal and 985 methaphorical adjective-noun pairs[10] and classify them. Gershman et al. (2014) further expand on this work using 64-dimensional vector-space word representations constructed by Faruqui and Dyer (2014) for classification. They report a state-of-the-art F-score of 85% with random decision forests, including also abstractness and imageability features (Wilson, 1988) and supersenses from WordNet, averaged across senses.

System	F1-score on test set
(Gershman et al., 2014)	85
WORDS	81.91±2.81
SUPER	**87.23**±2.36

Table 10: F1-score and a standard error on a provided test set for the adjective-noun metaphor prediction task Gershman et al. (2014). WORDS: word embeddings only, SUPER: multi-channel word embeddings with the supersense similarity and frequency vectors added

Since this setup is simpler than the sentence classification tasks, we use only a subset of our architecture, specifically the left half of Figure 3, i.e. our word embeddings, similarity vectors and supersense frequency vectors. Since there are only two words in each document, we leave out the LSTM layer. We merge the similarity and frequency layers by multiplication and concatenate the result to the word embedding convolution, feeding the output of the concatenation directly to the dense layer. Table 10 shows our results on a provided test set. Based on McNemar's test, there is a significant difference ($p < 0.01$) between our system based on words only and the one with supersenses.

7 Discussion

Unlike previous research on supersenses, our work is not based on a manually produced gold standard, but on an automatically annotated large corpus. While Scozzafava et al. (2015) report a high accuracy estimate of 77.8% on sense level, the performance and possible bias on tagged supersenses are yet to be evaluated. We are also aware that some of the previously proposed approaches for building word sense embeddings (Rothe and Schütze, 2015; Chen et al., 2014; Iacobacci et al., 2015) could be eventually extended to supersenses. We strongly encourage the authors to do so and perform a contrastive evaluation comparing these methods. Additionaly, a different level of granularity of the concepts, such as WordNet Domains (Magnini and Cavaglia, 2000) could be explored.

8 Conclusions and Future Work

We have presented a novel joint embedding set of words and supersenses, which provides a new insight into the word and supersense positions in the vector space. We demonstrated the utility of these embeddings for predicting supersenses and manifested that the supersense enrichment can lead to a significant improvement in a range of downstream classification tasks, using our embeddings in a neural network model. The outcomes of this work are available to the research community.[11]. In follow-up work, we aim to apply our embedding method on smaller, yet gold-standard corpora such as SemCor (Miller et al., 1994) and STREUSLE (Schneider and Smith, 2015) to examine the impact of the corpus choice in detail and extend the training data beyond WordNet vocabulary. Moreover, the coarse semantic categorization contained in supersenses was shown to be preserved in translation (Schneider et al., 2013), making them a perfect candidate for a multilingual adaptation of the vector space, e.g. extending Faruqui and Dyer (2014).

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant No. I/82806 and by the German Research Foundation under grant No. GU 798/14-1. Additional support was provided by the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus program under the reference 01-S12054 and by the German Institute for Educational Research (DIPF). We thank the anonymous reviewers for their input.

[10] http://www.cs.cmu.edu/~ytsvetko/metaphor/datasets.zip

[11] https://github.com/UKPLab/acl2016-supersense-embeddings

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 19–27. Association for Computational Linguistics.

Eneko Agirre, Kepa Bengoetxea, Koldo Gojenola, and Joakim Nivre. 2011. Improving dependency parsing with semantic classes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers-Volume 2*, pages 699–703. Association for Computational Linguistics.

Cem Akkaya, Janyce Wiebe, Alexander Conrad, and Rada Mihalcea. 2011. Improving the impact of subjectivity word sense disambiguation on contextual opinion analysis. In *Proceedings of the 15th Conference on Computational Natural Language Learning*, pages 87–96. Association for Computational Linguistics.

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Machine Learning and Knowledge Discovery in Databases*, pages 132–148. Springer.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *Conference on Artificial Intelligence*.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135.

Claudio Delli Bovi, Luis Espinosa Anke, and Roberto Navigli. 2015. Knowledge base unification via sense embeddings and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods on Natural Language Processing*, pages 726–736.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(1-47).

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing*, pages 1025–1035.

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 on Natural Language Processing*, pages 594–602. 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 on Natural Language Processing*, pages 168–175. Association for Computational Linguistics.

Hal Daumé III, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Machine learning*, 75(3):297–325.

Li Dong, Furu Wei, Shujie Liu, Ming Zhou, and Ke Xu. 2015. A statistical parsing framework for sentiment classification. *Computational Linguistics*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, April. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint*, arXiv:1605.02276.

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

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816.

Anatole Gershman, Yulia Tsvetkov, Leonid Boytsov, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd annual meeting on Association for Computational Linguistics*.

Josu Goikoetxea, Aitor Soroa, Eneko Agirre, and Basque Country Donostia. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 1434–1439.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. thesis, Carnegie Mellon University.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 95–105.

Rubén Izquierdo, Armando Suárez, and German Rigau. 2009. An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 389–397. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, Héctor Martınez Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of Twitter. *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, pages 1–11.

Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Massimiliano Ciaramita Jordi Atserias, Hugo Zaragoza and Giuseppe Attardi. 2008. Semantically annotated snapshot of the english wikipedia. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Egoitz Laparra and German Rigau. 2013. Impar: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1180–1189.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal, September. Association for Computational Linguistics.

Jiwei Li, Thang Luong, Dan Jurafsky, and Eduard Hovy. 2015. When are tree structures necessary for deep learning of representations? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2304–2314, Lisbon, Portugal, September. Association for Computational Linguistics.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating subject field codes into WordNet. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Mónica Marrero, Sonia Sánchez-Cuadrado, Jorge Morato Lara, and George Andreadakis. 2009. Evaluation of named entity extraction systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 746–751.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, pages 240–243. Association for Computational Linguistics.

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

George A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Alok Ranjan Pal and Diganta Saha. 2015. Word sense disambiguation: a survey. *arXiv preprint arXiv:1508.01346*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings*

of the 52nd Annual Meeting of the Association for Computational Linguistics, page 271. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing*, volume 14, pages 1532–1543.

Davide Picca, Alfio Massimiliano Gliozzo, and Massimiliano Ciaramita. 2008. Supersense tagger for italian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Citeseer.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of pos taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 968–973, Doha, Qatar, October. Association for Computational Linguistics.

Likun Qiu, Yunfang Wu, and Yanqiu Shao. 2011. Combining contextual and structural information for supersense tagging of Chinese unknown words. In *Computational Linguistics and Intelligent Text Processing*, pages 15–28. Springer.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*, pages 82–94.

Samuel Reese, Gemma Boleda Torrent, Montserrat Cuadros Oller, Lluís Padró, and German Rigau Claramunt. 2010. Word-sense disambiguated multilingual Wikipedia corpus. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and George A Vouros. 2009. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pages 370–375.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 Conference on Empirical Methods on Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China, July. Association for Computational Linguistics.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October.

Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 2011. Piggyback: Using search engines for robust cross-domain named entity recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 965–975. Association for Computational Linguistics.

Tara N Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4580–4584. IEEE.

Nathan Schneider and Noah A Smith. 2015. A corpus and model integrating multiword expressions and supersenses.

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: Short Papers-Volume 2*, pages 253–258. Association for Computational Linguistics.

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Olflazer, and Noah A Smith. 2013. Supersense tagging for arabic: the mt-in-the-middle attack. Association for Computational Linguistics.

Federico Scozzafava, Alessandro Raganato, Andrea Moro, and Roberto Navigli. 2015. Automatic identification and disambiguation of concepts and named entities in the multilingual wikipedia. In *AI* IA 2015 Advances in Artificial Intelligence*, pages 357–366. Springer.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013. Learning semantic textual similarity with structural representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 714–718.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods on Natural Language Processing*, pages 151–161. Association for Computational Linguistics.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Chiraag Sumanth and Diana Inkpen. 2015. How much does word sense disambiguation help in sentiment analysis of micropost data? In *6TH Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015)*, page 115.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383.

Yulia Tsvetkov, Anatole Gershman, and Elena Mukomel. 2013. Cross-lingual metaphor detection using common semantic features. In *The First Workshop on Metaphor in NLP*, page 45.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting english adjective senses with supersenses.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods on Natural Language Processing*. Association for Computational Linguistics.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85.

Sida Wang and Christopher Manning. 2013. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning*, pages 118–126.

Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 545–550.

Ye Zhang, Stephen Roller, and Byron Wallace. 2016. MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. *arXiv preprint arXiv:1603.00968*.