Exploitation of Co-reference in Distributional Semantics

Dominik Schlechtweg
University of Stuttgart
Stuttgart, Germany
dominik.schlechtweg@ling.uni-stuttgart.de

Abstract

The aim of distributional semantics is to model the similarity of the meaning of words via the words they occur with. Thereby, it relies on the distributional hypothesis implying that similar words have similar contexts. Deducing meaning from the distribution of words is interesting as it can be done automatically on large amounts of freely available raw text. It is because of this convenience that most current state-of-the-art-models of distributional semantics operate on raw text, although there have been successful attempts to integrate other kinds of—e.g., syntactic—information to improve distributional semantic models. In contrast, less attention has been paid to semantic information in the research community. One reason for this is that the extraction of semantic information from raw text is a complex, elaborate matter and in great parts not yet satisfyingly solved. Recently, however, there have been successful attempts to integrate a certain kind of semantic information, i.e., co-reference. Two basically different kinds of information contributed by co-reference with respect to the distribution of words will be identified. We will then focus on one of these and examine its general potential to improve distributional semantic models as well as certain more specific hypotheses.

Keywords: Distributional Semantics, Co-reference, Vector Space Models

1. Introduction

The original aim of distributional semantics is to model the similarity of the meaning—the semantics—of words. The basic assumption underlying this approach is that the semantic similarity of two words is a function of their contexts. That is, in other words, the meaning of words can be inferred from the frequencies of the words they immediately occur with and this can happen in such a way that the degree of similarity of those meanings can be measured. Semantic similarity is a key concept in the modeling of language and thus in computational linguistics. It is crucial in a variety of linguistic applications influencing our everyday life such as search engines.

In distributional semantics we represent the meaning of a word by a vector. This vector is an abstraction over the contexts in which we find the particular word in question. Here lies the crux of the matter: known algorithms of distributional semantics consider only those contexts as relevant to the meaning of a target word which are found as contexts of the *word*—the particular combination of letters, the string—in question. However, there are other, particularly definable, contexts which encode some of the meaning of a target word. Consider the following text example:

> When Cesar took on the case of {{Fella}, {an adorable Jack Russell/Italian Greyhound mix}}, {the little dog}'s antics were about to get {his} owner slapped with an eviction notice. At the apartment complex where {Fella} resided, {he} barked nonstop the entire time {his} adoptive mom was at work, ceasing only once she came home at night.[1]

Building up a vector representation for the meaning of *dog*, a standard algorithm of distributional semantics would browse through this text snippet, find the word *dog* just once, include the information from the immediate context (whose size would be defined previously) of this instance of *dog* into the vector representation and proceed. However, in the above it seems as though the contexts surrounding the noun phrases that refer to the same referent as the noun phrase which includes the word *dog*—such as *Fella*, *his* and *he*—are equally suited for contributing to the vector representation of the meaning of *dog*. Actually, the entire passage above is about a dog.

Now, if we want distributional models to use this information, we must make it explicit on the distributional level. Consider the following paragraph where we try to make co-reference information, which is implicit in the above text snippet, explicit:

> When Cesar took on the case of {Fella / an adorable... / the little dog / his / he}, {Fella / an adorable... / the little dog / his / he}'s antics were about to get {Fella / an adorable... / the little dog / his / he} owner slapped with an eviction notice. At the apartment complex where {Fella / an adorable... / the little dog / his / he} resided, {Fella / an adorable... / the little dog / his / he} barked nonstop the entire time {Fella / an adorable... / the little dog / his / he} adoptive mom was at work, ceasing only once she came home at night.

We may explicate it by finding co-referent noun phrases and telling the model at every spot a referent is picked up, e.g., by *he*, by which words it is elsewhere referred to (in all of the co-referent phrases), e.g., by *Fella* and *the little dog*. These alternative words used to refer to the same entity can then be put in the context of the current word, *he*. By this, the model gets access to the previously latent information, to the *latent contexts*.

Recently, there has already been an attempt to integrate co-reference information into models of distributional

[1] Cesar Millan. Cesar's Way. <http://www.cesarsway.com/dogbehavior/barking/What-Your-Dogs-Bark-is-Telling-You>. Last checked on March 10, 2016.

semantics (Schütze and Adel, 2014). Yet, the authors use a different kind of information than the one presented above. While we use *orthogonal* co-reference information, i.e., the contexts of co-referent mentions, Schütze and Adel use *linear* co-reference information, i.e., which mentions are actually co-referent, and consider co-referent mentions as mutual contexts.

We will compare standard distributional semantic models to models incorporating the above-described distributionally disguised co-reference information. While additional information gained through new contexts—what we will call here Orthogonal Context Enrichment (OCE)—should help the models in general, it could be particularly helpful for models trained on small data sets, since here the relative enrichment is higher than with bigger training set. The same rationale also applies for rare words: if there are not enough contexts for a word in a training set, then additional contexts gained through OCE may have a stronger impact on the learning of the meaning of the word than for a very frequent word. Also, OCE is expected to have a particular impact on learning the meaning of nouns, because co-reference is a relation holding between noun phrases, and especially proper names, which are very frequent in co-reference chains.

Apart from the general impact OCE may have on distributional semantic models, there are certain applications where we may imagine a particular benefit. OCE is, presumably, most helpful when raw text training data is limited, since here we cannot simply gather new contexts by scaling up the amount of training data. Hence, for tasks where we need to build many vectors from many small data points (instead of building one vector from large amounts of training data) we expect a particular benefit from OCE, since for every data point training data is limited. Such tasks occur, e.g., in word sense disambiguation, information retrieval, named-entity recognition or classification tasks such as spam detection. This makes OCE a widely applicable mechanism.

2. Procedure

In order to give distributional semantic models access to the above-mentioned latent contexts we first need a co-reference resolution for a corpus. For this we build on the *Annotated English Gigaword v.5 corpus* (Gigaword) annotated with syntactic and discourse structure and providing a co-reference resolution with quality of "current state of the art" (Napoles et al., 2012). We work with a subpart of around 50% of the size of the whole corpus encompassing approximately 100 million sentences and 1.9 billion tokens after preprocessing.

We then build a computational algorithm `headSub` replacing pronouns in the corpus with the head noun of the representative (most informative) element inside its co-reference chain, i.e., if the referent of a noun phrase n in the corpus is re-referred to via a pronoun p, then `headSub` replaces p with the syntactic head of n. By this we aim at enriching the corpus with more distributional information, as described above.[2] As an example the reader may consider the piece of text in (1).

(1) Fella chases a squirrel, since he wants to eat it.

The co-reference resolution for (1) shall be $\{(Fella, he), (a\ \ squirrel, it)\}$ with the first elements of the chains being the representative element respectively. Now `headSub` will produce the following output text for (1):

(2) fella chases a squirrel since fella wants to eat squirrel

We also experiment with different versions of `headSub`, e.g., we only insert proper nouns ($\texttt{headSub}_{\texttt{PN}}$) or lexical nouns ($\texttt{headSub}_{\texttt{Non-PN}}$).

After this, we train a distributional semantic model on the enriched text (corresponding to (2) in the example above) and for comparison also on the original raw text (corresponding to (1)). For training we use the Skip-gram model from the *word2vec* toolkit (Mikolov et al., 2013a; Mikolov et al., 2013b), which was found to be superior to standard count models (Baroni et al., 2014). Since the relative performance of the different models was found to vary with the variation of these training parameters we vary the maximal window size considered as the context of a token and the minimum count of words considered during training in order to get a broader picture of the impact of OCE.

Finally, the resulting vector spaces will be evaluated with respect to their capturing of semantic (attributional) similarity.

3. Evaluation

We evaluate the quality of the vector spaces on a variety of data sets using two basically different ways of evaluation: (i) human similarity judgments of word pairs and (ii) analogies.

The word similarity judgments consist of pairs of words rated by human informants for their semantic similarity. Evaluation here means determining the Spearman's rank correlation coefficient between all similarity judgments for word pairs inside a test set and the cosine similarities of the respective word vectors. We will evaluate the vector spaces on a variety of human similarity judgment test sets including standard benchmarks such as WordSim353 (Finkelstein et al., 2002; Agirre et al., 2009) (which will allow us to distinguish between similarity and relatedness)[3], SimLex-999 (Hill et al., 2014) (which will allow us to distinguish between different parts of speech, i.e., nouns, adjectives and verbs) and MEN (Bruni et al., 2014) plus a data set containing mainly rare words, which we will call Rare (Luong et al., 2013), and a data set containing many proper nouns, which we will call MTurk (Radinsky et al., 2011). These test sets are found to have a very different constitution concerning the classes of words they contain. While

[2] The reader may note that a similar idea was already applied for sentiment analysis in (Pontiveros, 2012).

[3] Note that we will exclude the word pairs containing proper nouns from WordSim353, since we do not want effects concerning proper nouns to intervene with effects concerning the distinction between similarity and relatedness.

some contain mainly nouns, others contain mainly adjectives or verbs. Also, the use of proper nouns strongly varies; some do not even contain any proper nouns. As this study also indicates, this varying constitution of test sets may lead to very different results testing a model on them.

The second evaluation procedure follows the idea that there are different kinds of similarities between words (Mikolov et al., 2013a). Evaluation here consists of checking whether a respective vector space captures certain relations via comparing the distances of word pairs with the same relation. For example, the distance between *great* and *greatest* should be the same as between *smart* and *smartest*. Analogies provide a convenient way of evaluation, since it is easy to construe new sets testing very different relations involving many different kinds of words. We exploit this fact by construing three new analogy test sets which we provide as an additional resource to this study: one set consisting of word2vec's analogy questions (Mikolov et al., 2013a) reconstructed with rare words and two sets for proper nouns, one based on words and the other based on phrases (Dominik Schlechtweg, 2016). These new resources will help us in measuring the quality of a vector space with respect to its ability to capture the similarity properties of rare words and proper nouns, on which we assumed OCE to have a particular impact. Additionally, we evaluate the vector spaces on word2vec's analogy questions.

3.1. Rare Words

The test set for rare words—exemplified in Table 1—was construed on the basis of word2vec's analogy questions. We first excluded the *(Common) Capital city* task and the *Man-Woman* (*family*) task from consideration, since the former necessarily contains frequent words, while for the latter we could not imagine enough rare words and did not find a way to search for it in a corpus. For other tasks this was possible. For the *City-in-state* task we combined English names of Chinese districts with names of their capitals (instead of American states and their capitals in the original file). For the remaining tasks we scanned approximately 1.3 million sentences from the *Annotated English Gigaword v.5 corpus*. For the *Currency* task, for instance, we searched for words with the NER-tag *MONEY*, or for the *Adjective to adverb* task we searched for words with the respective POS-tag. Then we worked manually through the rarest words of the respective category (those with frequency below 10) and selected words that seemed well-suited because they were of the specific type required for the task. The selected item, say *ghastly*, was then combined with the related element according to the target task; for the comparative task this would be *ghastlier*. This pair was then, in turn, combined with all of the pairs from the same relation in the word2vec *questions-words* file, and vice versa. By this procedure we always combine one pair which was extracted by us with one pair from the *questions-words* file, i.e., one rare pair with a more frequent one. In this way we want to avoid the "adding up" of the rareness of the word pairs which otherwise may lead to an extreme drop of performance on the tasks. Moreover, this leads to the effect that we get a large number of questions. In this way we get a total of 29,150 questions, nearly 10,000 more than in the *questions-words* file.[4] We provide this data set as an additional resource to this paper, since it might be a complementary utility to the word2vec *questions-words* file for further research.

Type of relationship	Word Pair I	Word Pair II
Capital city	Algiers Algeria	Vaduz Liechtenstein
Currency	India rupee	Swaziland emalangeni
Chinese city-in-state	Nanning Guangxi	Wuhan Hubei
Adjective to adverb	quick quickly	surreptitious surreptitiously
Opposite	rational irrational	vaccinated unvaccinated
Comparative	bad worse	ghastly ghastlier
Superlative	bad worst	thorny thorniest
Present participle	code coding	nullify nullifying
Nationality adjective	Israel Israeli	Barbados Barbadian
Past tense	dancing danced	filching filched
Plural nouns	cow cows	raccoon raccoons
Plural verbs	slow slows	peg pegs

Table 1: Structure of the *questions-words-rare* file: word2vec's analogy questions reconstructed with rare words.

3.2. Proper Nouns

In order to evaluate the models specifically with respect to the similarity properties of proper nouns we chose four relations involving proper nouns: the *leader-country* relation and the *person-sex* relation shall measure similarity properties of names of humans, while the relations *building-city* and *river-country* shall measure similarity properties of names of things. The word pairs contain mostly names of well-known entities, such as former or present state leaders, nations, presently famous or historically important people, buildings, cities and rivers. The structure of the *questions-words-proper-nouns* file containing a total of 2,746 questions is depicted in Table 2. The *questions-phrases-proper-nouns* file based on phrases has the same structure. Both files are provided as an additional resource to this paper.[5]

Type of relationship	Word Pair I	Word Pair II
leader-country	Obama USA	Putin Russia
person-sex	Cleopatra woman	Einstein man
building-city	Reichstag Berlin	Kremlin Moscow
river-country	Nile Egypt	Rhine Germany

Table 2: Structure of the *questions-words-proper-nouns* file.

4. Results

4.1. A Preliminary Study

A preliminary study showed varying results across the different tasks we evaluated on. Yet, on analogies one OCE-model, i.e., that is trained on an enriched text, outperformed the other OCE-models on most evaluation tasks and—at

[4]Note that by allowing two rare word pairs in the same question we could further increase this number without any additional effort.

[5]The reader may note that there is a bias towards male entities in the files. However, in an updated version this bias shall be eliminated.

least on analogies—also the baseline model, trained on the original raw text. Surprisingly, this was $\texttt{headSub}_{\texttt{PN}}$, inserting only proper nouns for pronouns, which we initially did not expect to contribute with so much distributional information.[6] For this, $\texttt{headSub}_{\texttt{PN}}$ was chosen for a deeper analysis. This is not to say that the other OCE-models (headSub and $\texttt{headSub}_{\texttt{Non-PN}}$) are not expected to contribute to the learning of the meaning of certain kinds of words. But, with the present means at hand, focusing on one model which showed the clearest results seemed to be the best option. The reader may note, however, that by this restriction and also by the specific restrictions resulting from the mechanism of headSub and its derivatives we presumably only exploit a small share of the full potential of OCE.

4.2. A Deeper Analysis: Inserting Proper Nouns

The reader may consider Table 3 for the results of training skip-gram on the output of $\texttt{headSub}_{\texttt{PN}}$ and raw text (the baseline) respectively with varying training parameters. The models were trained 10 times each. In every iteration the performance on the different test sets was computed at the end of the training period. Here we present the average performance over all iterations. An independent two-sample t-test was performed for each set of training parameters between the results of the models trained on the output of $\texttt{headSub}_{\texttt{PN}}$ and the results of the baseline models in order to assure statistical significance of the results.[7] The resulting p-value is given for each test set. Performance values are marked boldly where for a certain combination of training parameters the models trained on one of the two texts ($\texttt{headSub}_{\texttt{PN}}$ or baseline) outperformed the models trained on the other *and* the difference is statistically significant, i.e., the p-value is below 0.01.

4.2.1. General Observations

Generally, we do not find very strong performance differences. The strongest are around 2%. Yet, the first striking observation considering Table 3 is the different performance of the models on similarity judgments and analogies: while the models trained on raw text (baseline-models) have significant advantages on many similarity judgment test sets, the situation looks the other way round on analogies where the models trained on the output of $\texttt{headSub}_{\texttt{PN}}$ ($\texttt{headSub}_{\texttt{PN}}$-models) have significant advantages. Why is that?

4.2.2. Similarity vs. Relatedness

We may find the reason for the different performances on the two evaluation methodologies in the distinction between similarity and relatedness. While certain similarity judgment test sets particularly aim at measuring similarity and not relatedness (SimLex-999) or distinguish between similarity and relatedness (WordSim353), the analogy test sets—at least in the particular form at hand here—seem to be more suited to measure relatedness than similarity. Two words are *related* if they "are associated but not [necessarily] actually similar (*Freud*, *psychology*)" (Hill et al., 2014) or if they "are connected by broader semantic relations" (Bruni et al., 2014). The latter is exactly the way in which the analogy test sets were construed, i.e., words with the same relation are checked for equal distances in vector space. (The reader may note that these relations have very different natures.) Hence, we can expect the analogy test sets used here to be a better measure for relatedness rather than the more narrow notion of similarity. This is also supported by the performance of the models on WordSim353. Though not statistically significant yet, we observe advantages of the baseline-models on the similarity measure, while we observe advantages of $\texttt{headSub}_{\texttt{PN}}$-models on the relatedness measure. If the analogy test sets used here are more suited to measure relatedness and the similarity judgment test sets rather measure similarity, the different performances on these two methodologies are explained by the advantages of the $\texttt{headSub}_{\texttt{PN}}$-models in capturing relatedness.[8]

4.2.3. Nouns

The advantage of $\texttt{headSub}_{\texttt{PN}}$-models for one set of parameters on the noun subset of SimLex-999 indicates that—under certain circumstances—OCE might be helpful for learning the meaning of nouns. As we already mentioned above, this would not be surprising in general, since co-reference—and thus the mechanism of $\texttt{headSub}_{\texttt{PN}}$—mainly involves insertion of nouns. However, for the particular model used here, i.e., $\texttt{headSub}_{\texttt{PN}}$, this effect is indeed surprising, since it only inserts proper nouns, but these are not part of the SimLex-999 noun-subset.

4.2.4. Verbs

The results for verbs on similarity judgments are clear: with all training parameter sets we have significant advantages of the baseline-models. On analogies, though, for tasks involving verbs such as *Past tense* or *Present participle* the baseline is significantly outperformed for different parameter sets and also on rare words.

4.2.5. Adjectives

While for adjectives we have no significant differences on similarity judgments, we do find significant differences on the *Adjective to adverb* task for different parameter sets and for frequent as well as for rare words in favor of the $\texttt{headSub}_{\texttt{PN}}$-models. However, for the *Nationality adjective* task the baseline still significantly outperforms the OCE-model on one set of training parameters.

4.2.6. Rare Words

On Rare we do not find significant differences. However, on the word2vec analogy questions reconstructed with rare

[6] It is not yet clear what the reason for this effect is. The fact that the quality of the co-reference resolution for proper nouns is better than for other kinds of words may have an influence.

[7] A normal distribution of the results was assumed.

[8] However, it is not clear why the baseline has clear advantages on MEN, of which the authors explicitly claim to measure relatedness. Note, however, that this does not mean that the test set does not measure similarity at all. It rather means that it measures both, since relatedness covers similarity. Thus, a possible explanation would be that there is still a focus on genuine similarity judgments in this test set.

Test set	min_{50}, window_5			min_{25}, window_5			min_{50}, window_3		
	$\texttt{headSub}_{\texttt{PN}}$	baseline	p-value	$\texttt{headSub}_{\texttt{PN}}$	baseline	p-value	$\texttt{headSub}_{\texttt{PN}}$	baseline	p-value
SIMILARITY JUDGMENTS									
WordSim353									
similarity	70.47	71.20	0.012	70.79	71.30	0.020	71.89	72.24	0.159
relatedness	59.65	59.17	0.044	59.55	59.19	0.104	59.32	58.73	0.272
SimLex-999	41.43	**42.14**	< 0.01	41.66	**42.28**	< 0.01	43.17	43.40	0.036
nouns	43.04	43.08	0.705	43.39	43.48	0.362	**44.17**	43.89	< 0.01
adjectives	57.62	58.21	0.021	57.87	58.44	0.057	59.21	59.02	0.592
verbs	26.91	**29.59**	< 0.01	27.02	**29.12**	< 0.01	30.31	**32.06**	< 0.01
MEN	72.28	**72.60**	< 0.01	72.36	**72.68**	< 0.01	72.89	**73.08**	< 0.01
Rare	51.87	51.70	0.158	48.82	48.78	0.754	52.03	52.03	0.954
MTurk	68.74	68.71	0.896	68.73	68.13	0.060	**68.22**	67.46	< 0.01
proper nouns	69.23	69.23	0.994	68.73	67.38	0.065	69.76	68.11	0.011
ANALOGIES									
Word2vec	**67.39**	67.04	< 0.01	**66.49**	66.11	< 0.01	68.14	67.98	0.225
Common capital city	90.61	89.64	0.043	**89.64**	88.22	< 0.01	90.97	90.55	0.224
All capital cities	90.16	89.89	0.213	**88.80**	87.97	< 0.01	89.96	89.85	0.561
Currency	15.75	16.10	0.172	d. c.	d. c.	–	17.64	17.79	0.746
City-in-state	57.13	56.55	0.05	55.81	55.91	0.748	58.85	59.31	0.165
Man-Woman	72.93	73.95	0.097	71.13	**73.44**	< 0.01	73.60	74.72	0.127
Adjective to adverb	**25.39**	24.13	< 0.01	**24.33**	23.11	< 0.01	22.94	21.98	0.014
Opposite	34.32	34.40	0.883	33.65	33.13	0.185	35.64	35.69	0.925
Comparative	86.59	86.79	0.559	86.34	85.92	0.174	88.00	87.76	0.428
Superlative	57.18	56.48	0.261	56.00	54.16	0.023	59.42	58.49	0.340
Present Participle	61.73	61.46	0.598	**63.10**	61.35	< 0.01	62.40	62.56	0.775
Nationality adjective	87.86	87.78	0.573	87.81	**88.27**	< 0.01	89.17	89.27	0.599
Past tense	62.92	62.00	0.036	**62.96**	61.66	< 0.01	64.00	64.74	0.030
Plural nouns	68.53	68.46	0.872	66.98	66.78	0.718	69.11	69.07	0.925
Plural verbs	49.31	48.37	0.136	48.77	47.70	0.022	50.07	49.41	0.120
Word2vec rare	36.96	36.77	0.217	**35.14**	34.54	< 0.01	37.69	37.41	0.054
Capital city	76.49	76.06	0.368	75.01	74.03	0.303	75.56	75.83	0.481
Currency	12.63	12.87	0.279	d. c.	d. c.	–	14.75	14.71	0.895
Chinese city-in-state	98.00	97.70	0.254	98.28	98.12	0.614	96.12	95.50	0.253
Adjective to adverb	**18.02**	17.33	< 0.01	15.32	15.17	0.478	16.73	16.80	0.770
Opposite	13.50	13.66	0.722	11.76	11.49	0.282	13.94	13.40	0.189
Comparative	65.24	65.45	0.801	65.41	65.38	0.969	68.32	69.50	0.104
Superlative	41.49	42.20	0.562	42.43	41.37	0.426	48.88	47.38	0.115
Present Participle	42.62	42.60	0.972	41.21	40.22	0.018	43.43	43.71	0.519
Nationality adjective	75.86	75.87	0.991	77.10	75.27	0.013	79.34	77.93	0.029
Past tense	37.37	36.68	0.075	**36.51**	34.70	< 0.01	38.34	37.31	0.012
Plural nouns	41.10	41.56	0.372	38.59	38.29	0.528	41.82	41.29	0.144
Plural verbs	30.04	29.66	0.295	28.96	29.01	0.942	31.19	31.05	0.737
Proper nouns	24.11	23.31	0.053	22.60	22.38	0.426	25.72	24.75	0.061
leader-country	**22.65**	20.99	< 0.01	**22.21**	20.33	< 0.01	**26.14**	24.30	< 0.01
person-sex	err.	err.	–	err.	err.	–	err.	err.	–
building-city	30.69	30.97	0.824	27.64	28.61	0.392	29.58	30.70	0.272
river-country	23.68	23.74	0.945	21.52	22.91	0.044	23.57	23.08	0.546

Table 3: Performance of skip-gram model trained on output of $\texttt{headSub}_{\texttt{PN}}$ and raw text (the baseline) with varying training parameters. For analogies accuracy values are given, while for similarity judgments we give the Spearman's rank correlation coefficient (multiplied by 100 for better comparison with the accuracy values). Tasks where models show highly different coverage of the data are excluded (marked by "d. c."). The *person-sex* task was subsequently excluded because errors in the test set led to biased results (marked by "err.").

words (Word2vec rare) the baseline is significantly outperformed for one parameter set, while the performance for the other parameter sets confirms this tendency. The overall advantage of the $\texttt{headSub}_{\texttt{PN}}$-models on Word2vec rare is comparable to the advantage on the original word2vec questions. A particular improvement for rare words with OCE can thus not be confirmed here.

4.2.7. Proper Nouns

As we use an OCE-model inserting only proper nouns we would expect a particular effect for proper nouns. Also, because animate entities are more likely to be re-referred to we would expect a stronger effect for proper names of human entities. There are indeed significant advantages of the $\texttt{headSub}_{\texttt{PN}}$-models when it comes to test tasks involving proper names of human entities.[9] This is indicated here by their performance on *leader-country*: the baseline is significantly outperformed for all training parameters.

Further, for most of the tasks in the word2vec analogy set involving proper nouns, such as *Common capital city* and *All capital cities* the baseline is outperformed significantly for one set of training parameters, which is supported by the performances with the other parameters.

Also, for similarity judgments these observations are confirmed: on MTurk, containing a comparably high number of proper nouns, the baseline is outperformed significantly for one parameter set, where this tendency is confirmed for the other sets. Further, for the subset of proper nouns from MTurk there is—though not clearly significant—a tendency towards advantages of the $\texttt{headSub}_{\texttt{PN}}$-models.

4.2.8. Pronouns

The only task containing pronouns is *Man-Woman*. Since $\texttt{headSub}_{\texttt{PN}}$ deletes many pronouns we may expect a particular effect here. This is indeed the case. For one set of training parameters the baseline-models significantly outperform the $\texttt{headSub}_{\texttt{PN}}$-models, which is also supported by the performances on the other parameter sets.

5. Conclusion

In the above we found that one particular way of carrying out OCE—i.e, by replacing pronouns with the syntactic head of the proper nouns they are co-referent with—de facto improves the performance of distributional models on a wide range of analogy tasks and on certain test sets of similarity judgments. We find the clearest results for proper nouns (more specifically, for proper names of human entities), which was what we expected, since only proper nouns were inserted by $\texttt{headSub}_{\texttt{PN}}$. Yet, also for nouns in general, adjectives and verbs the findings indicate that OCE may have a potential to improve the learning of their semantic similarity—or perhaps relatedness—properties. However, we also found significant disadvantages of the OCE-model used here, especially on test sets of similarity judgments. The initial hypothesis that OCE will help for learning the meaning of rare words could not be confirmed. Whether the different performances on analogies and similarity judgments are indeed due to the distinction between relatedness and similarity has to be examined more deeply.

The reader may, however, note that a major downside of OCE is its reliance on co-reference resolution which makes it a computationally costly, supervised and language dependent approach in contrast to standard models of distributional semantics. Also, it is strongly dependent on the quality of co-reference resolution, which is—in the best case—around 60% (F1) for present co-reference resolution algorithms (Lee et al., 2011).

In the end, also, the results obtained above have to be checked, not only because they vary across tasks, but also because the operation carried out by `headSub` and its derivatives may trigger certain side effects that also may have an influence on the performance of the resulting vector space models.[10] In order to exclude these factors and in order to exploit the full potential of OCE co-reference information shall be integrated directly into the training process of a standard count model. That is, we will retreat from first integrating co-reference information into raw text and then performing training. Instead, we will build a standard count model of distributional semantics sensitive to co-reference information by directly accessing the context of co-referent mentions when encountering a mention which is part of a co-reference chain.

Further, a qualitative analysis of the resulting vector spaces has to be carried out in order to explain how the different performances caused by OCE come about.[11] That is also the question whether we can regard OCE as yielding just more data of the same kind as the linear distribution of words or whether we may gather new—otherwise possibly rare—kinds of information. Also, we may evaluate the effect of OCE on smaller data sets. Above that, the best way to make orthogonal information distributionally explicit has to be examined, i.e., we have to find out which are the sets of words to replace and to insert which yield the best results for a certain task; recall that the OCE-model presented here is restricted in many ways and presumably only ex-

[9] The results on the *person-sex* task had to be excluded because they were biased due to errors in the test set. Yet, in previous experiments certain OCE-models constantly outperformed baseline models on this task, in particular with respect to feminine entities.

[10] Some of these side effects are *window effects*. By substituting one or more tokens we may "push out" or "pull in" other tokens from the training window. Pushing out may happen for instance when we insert more than one token for another token. Pulling in may happen when there was carried out a substitution in the context of a token considered during training but the inserted word was deleted during training, for instance by subsampling or because of the minimum word count.

[11] If we assume that the orthogonal distribution (in contexts of co-referent words) of words is generally (on average) the same as their linear distribution, then the performance improvements are just explainable by the fact that we gain more data (more contexts), which convey no information which could not—in principle—be gained by considering more linear contexts. Sure, the rarer the word, the more linear contexts we would have to consider (on average) in order to find the information we search for. Whether this assumption is indeed valid has to be examined in the future. Only if we assumed that certain words or constructions tended to co-occur with pronouns rather than with co-referent richer descriptions, we could say that there is a new—complementary—type of information gained through OCE explaining differences in performance.

ploits a small share of the potential OCE has. Finally, OCE should be evaluated directly with relevant applications as, e.g., information retrieval, classification tasks or sense disambiguation.

6. Acknowledgments

I thank Sandra Herrmann, Sascha Schlechtweg, Stefanie Eckmann, Tatjana Schlechtweg and Veronika Vasileva for intensive last-minute help.

7. Bibliographical References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, Maryland, USA.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Hill, F., Reichart, R., and Korhonen, A. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA. Association for Computational Linguistics.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *CoNLL*, Sofia, Bulgaria.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Napoles, C., Gormley, M., and Durme, B. V. (2012). Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100.

Pontiveros, B. B. F. (2012). Opinion mining from a large corpora of natural language reviews. Master's thesis, LSI, UPC.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, Hyderabad, India.

Schütze, H. and Adel, H. (2014). Using mined coreference chains as a resource for a semantic task. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1447–1452, Doha, Qatar. ACL.

8. Language Resource References

Dominik Schlechtweg. (2016). *Analogy questions involving rare words and proper nouns for evaluation of vector space models of distributional semantics*. Dominik Schlechtweg, distributed via ELRA, 1.0.