Sense and Similarity: A Study of Sense-level Similarity Measures

Nicolai Erbs[†], Iryna Gurevych[†‡] and Torsten Zesch[§]

† UKP Lab, Technische Universität Darmstadt
‡ Information Center for Education, DIPF, Frankfurt
§ Language Technology Lab, University of Duisburg-Essen
http://www.ukp.tu-darmstadt.de

Abstract

In this paper, we investigate the difference between word and sense similarity measures and present means to convert a state-of-the-art word similarity measure into a sense similarity measure. In order to evaluate the new measure, we create a special sense similarity dataset and re-rate an existing word similarity dataset using two different sense inventories from WordNet and Wikipedia. We discover that word-level measures were not able to differentiate between different senses of one word, while sense-level measures actually increase correlation when shifting to sense similarities. Sense-level similarity measures improve when evaluated with a re-rated sense-aware gold standard, while correlation with word-level similarity measures decreases.

1 Introduction

Measuring similarity between words is a very important task within NLP with applications in tasks such as word sense disambiguation, information retrieval, and question answering. However, most of the existing approaches compute similarity on the word-level instead of the sense-level. Consequently, most evaluation datasets have so far been annotated on the word level, which is problematic as annotators might not know some infrequent senses and are influenced by the more probable senses. In this paper, we provide evidence that this process heavily influences the annotation process. For example, when people are presented the word pair *jaguar* - *gamepad* only few people know that *jaguar* is also the name of an Atari game console.[1] People rather know the more common senses of *jaguar*, i.e. the car brand or the animal. Thus, the word pair receives a low similarity score, while computational measures are not so easily fooled by popular senses. It is thus likely that existing evaluation datasets give a wrong picture of the true performance of similarity measures.

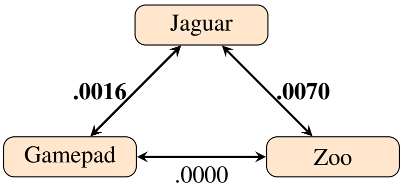


Figure 1: Similarity between words.

Thus, in this paper we investigate whether similarity should be measured on the sense level. We analyze state-of-the-art methods and describe how the word-based Explicit Semantic Analysis (ESA) measure (Gabrilovich and Markovitch, 2007) can be transformed into a sense-level measure. We create a sense similarity dataset, where senses are clearly defined and evaluate similarity measures with this novel dataset. We also re-annotate an existing word-level dataset on the sense level in order to study the impact of sense-level computation of similarity.

2 Word-level vs. Sense-level Similarity

Existing measures either compute similarity (i) on the word level or (ii) on the sense level. Similarity on the word level may cover any possible sense of the word, where on the sense level only the actual sense is considered. We use Wikipedia Link Mea-

[1] If you knew that it is a certain sign that you are getting old.

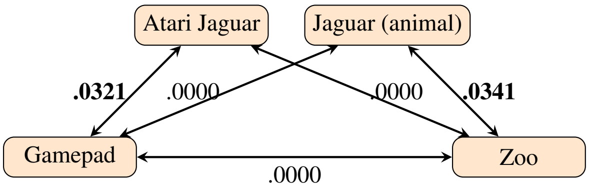


Figure 2: Similarity between senses.

sure (Milne, 2007) and Lin (Lin, 1998) as examples of sense-level similarity measures[2] and ESA as the prototypical word-level measure.[3]

The Lin measure is a widely used graph-based similarity measure from a family of similar approaches (Budanitsky and Hirst, 2006; Seco et al., 2004; Banerjee and Pedersen, 2002; Resnik, 1999; Jiang and Conrath, 1997; Grefenstette, 1992). It computes the similarity between two senses based on the information content (IC) of the lowest common subsumer (lcs) and both senses (see Formula 1).

$$\text{sim}_{\text{lin}} = \frac{2\ IC(lcs)}{IC(sense1) + IC(sense2)} \quad (1)$$

Another type of sense-level similarity measure is based on Wikipedia that can also be considered a sense inventory, similar to WordNet. Milne (2007) uses the link structure obtained from articles to count the number of shared incoming links of articles. Milne and Witten (2008) give a more efficient variation for computing similarity (see Formula 2) based on the number of links for each article, shared links $|A \cap B|$ and the total number of articles in Wikipedia $|W|$.

$$\text{sim}_{\text{LM}} = \frac{\log max(|A|,|B|) - \log|A \cap B|}{\log|W| - \log min(|A|,|B|)} \quad (2)$$

All sense-level similarity measures can be converted into a word similarity measure by computing the maximum similarity between all possible sense pairs. Formula 3 shows the heuristic, with S_n being the possible senses for word n, sim_w the word similarity, and sim_s the sense similarity.

$$\text{sim}_w(w_1, w_2) = \max_{s_1 \in S_1, s_2 \in S_2} \text{sim}_s(s1, s2) \quad (3)$$

Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) is a widely used word-level similarity measure based on Wikipedia as a background document collection. ESA constructs a n-dimensional space, where n is the number of articles in Wikipedia. A word is transformed in a vector with the length n. Values of the vector are determined by the term frequency in the corresponding dimension, i.e. in a certain Wikipedia article. The similarity of two words is then computed as the inner product (usually the cosine) of the two word vectors.

We now show how ESA can be adapted successfully to work on the sense-level, too.

2.1 DESA: Disambiguated ESA

In the standard definintion, ESA computes the term frequency based on the number of times a term—usually a word—appears in a document. In order to make it work on the sense level, we will need a large sense-disambiguated corpus. Such a corpus could be obtained by performing word sense disambiguating (Agirre and Edmonds, 2006; Navigli, 2009) on all words. However, as this is an error-prone task and we are more interested to showcase the overall principle, we rely on Wikipedia as an already manually disambiguated corpus. Wikipedia is a highly linked resource and articles can be considered as senses.[4] We extract all links from all articles, with the link target as the term. This approach is not restricted to Wikipedia, but can be applied to any resource containing connections between articles, such as Wiktionary (Meyer and Gurevych, 2012b). Another reason to select Wikipedia as a corpus is that it will allow us to directly compare similarity values with the Wikipedia Link Measure as described above.

After this more high-level introduction, we now focus on the mathematical foundation of ESA and disambiguated ESA (called ESA on senses). ESA and ESA on senses count the frequency of each term (or sense) in each document. Table 1 shows the corresponding term-document matrix for the example in Figure 1. The term *Jaguar* appears in all shown documents, but the term *Zoo* appears in the articles *Dublin Zoo* and *Wildlife Park*.[5] A manual analysis shows that *Jaguar* appears with different senses in the articles *D-pad*[6] and *Dublin Zoo*.

[2] We selected these measures because they are intuitive but still among the best performing measures.

[3] Hassan and Mihalcea (2011) classify these measures as corpus-based and knowledge-based.

[4] Wikipedia also contains pages with a list of possible senses called *disambiguation pages*, which we filter.

[5] In total it appears in 30 articles but we shown only few example articles.

[6] A D-pad is a directional pad for playing computer games.

Articles	Terms		
	Jaguar	*Gamepad*	*Zoo*
# articles	3,496	30	7,553
Dublin Zoo	1	0	25
Wildlife Park	1	0	3
D-pad	1	0	0
Gamepad	4	1	0
...	...	...	...

Table 1: Term-document-matrix for frequencies in a corpus if words are used as terms

Articles	Terms			
	Atari Jaguar	Gamepad	Jaguar (animal)	Zoo
# articles	156	86	578	925
Dublin Zoo	0	0	2	1
Wildlife Park	0	0	1	1
D-pad	1	1	0	0
Gamepad	1	0	0	0
...	...	...	...	...

Table 2: Term-document-matrix for frequencies in a corpus if senses are used as terms

By comparing the vectors without any modification, we see that the word pairs *Jaguar—Zoo* and *Jaguar—Gamepad* have vector entries for the same document, thus leading to a non-zero similarity. Vectors for the terms *Gamepad* and *Zoo* do not share any documents, thus leading to a similarity of zero.

Shifting from words to senses changes term frequencies in the term-document-matrix in Table 2. The word *Jaguar* is split in the senses *Atari Jaguar* and *Jaguar (animal)*. Overall, the term-document-matrix for the sense-based similarity shows lower frequencies, usually zero or one because in most cases one article does not link to another article or exactly once. Both senses of *Jaguar* do not appear in the same document, hence, their vectors are orthogonal. The vector for the term *Gamepad* differs from the vector for the same term in Table 1. This is due to two effects: (i) There is no link from the article *Gamepad* to itself, but the term is mentioned in the article and (ii) there exists a link from the article *D-pad* to *Gamepad*, but using another term.

The term-document-matrices in Table 1 and 2 show unmodified frequencies of the terms. When comparing two vectors, both are normalized in a prior step. Values can be normalized by the inverse logarithm of their document frequency. Term frequencies can also be normalized by weighting them with the inverse frequency of links pointing to an article (document or articles with many links pointing to them receive lower weights as documents with only few incoming links.) We normalize vector values with the inverse logarithm of article frequencies.

Besides comparing two vectors by measuring the angle between them (cosine), we also experiment with a language model variant. In the language model variant we calculate for both vectors the ratio of links they both share. The final similarity value is the average for both vectors. This is somewhat similar to the approach of Wikipedia Link Measure by Milne (2007). Both rely on Wikipedia links and are based on frequencies of these links. We show that—although, ESA and Link Measure seem to be very different—they both share a general idea and are identical with a certain configuration.

2.2 Relation to the Wikipedia Link Measure

Link Measure counts the number of incoming links to both articles and the number of shared links. In the originally presented formula by Milne (2007) the similarity is the cosine of vectors for incoming or outgoing links from both articles. Incoming links are also shown in term-document-matrices in Table 1 and 2, thus providing the same vector information. In Milne (2007), vector values are weighted by the frequency of each link normalized by the logarithmic inverse frequency of links pointing to the target. This is one of the earlier described normalization approaches. Thus, we argue that the Wikipedia Link Measure is a special case of our more general ESA on senses approach.

3 Annotation Study I: Rating Sense Similarity

We argue that human judgment of similarity between words is influenced by the most probable sense. We create a dataset with ambiguous terms and ask annotators to rank the similarity of senses and evaluate similarity measures with the novel dataset.

3.1 Constructing an Ambiguous Dataset

In this section, we discuss how an evaluation dataset should be constructed in order to correctly asses the similarity of two senses. Typically, evaluation datasets for word similarity are constructed by letting annotators rate the similarity between both words without specifying any senses for these words. It is common understanding that annotators judge the similarity of the combination of senses with the highest similarity.

We investigate this hypothesis by constructing a new dataset consisting of 105 ambiguous word pairs. Word pairs are constructed by adding one word with two clearly distinct senses and a second word, which has a high similarity to only one of the senses. We first ask two annotators[7] to rate the word pairs on a scale from 0 (not similar at all) to 4 (almost identical). In the second round, we ask the same annotators to rate 277 sense[8] pairs for these word pairs using the same scale.

The final dataset thus consists of two levels: (i) word similarity ratings and (ii) sense similarity ratings. The gold ratings are the averaged ratings of both annotators, resulting in an agreement[9] of .510 (Spearman: .598) for word ratings and .792 (Spearman: .806) for sense ratings.

Table 3 shows ratings of both annotators for two word pairs and ratings for all sense combinations. In the given example, the word *bass* has the senses of the fish, the instrument, and the sound. Annotators compare the words and senses to the words *Fish* and *Horn*, which appear only in one sense (most frequent sense) in the dataset.

The annotators' rankings contradict the assumption that the word similarity equals the similarity of the highest sense. Instead, the highest sense similarity rating is higher than the word similarity rating. This may be caused—among others—by two effects: (i) the correct sense is not known or not recalled, or (ii) the annotators (unconsciously) adjust their ratings to the probability of the sense. Although, the annotation manual stated that Wikipedia (the source of the senses) could be used to get informed about senses and that any sense for the words can be selected, we see both effects in the annotators' ratings. Both annotators rated the similarity between *Bass* and *Fish* as very low (1 and 2). However, when asked to rate the similarity between the sense *Bass (Fish)* and *Fish*, both annotators rated the similarity as high (4). Accordingly, for the word pair *Bass* and *Horn*, word similarity is low (1) while the highest sense frequency is medium to high (3 and 4).

3.2 Results & Discussion

We evaluated similarity measures with the previously created new dataset. Table 4 shows correlations of similarity measures with human ratings. We divide the table into measures computing similarity on word level and on sense level. ESA works entirely on a word level, Lin (WordNet) uses WordNet as a sense inventory, which means that senses differ across sense inventories.[10] ESA on senses and Wikipedia Link Measure (WLM) compute similarity on a sense-level, however, similarity on a word-level is computed by taking the maximum similarity of all possible sense pairs.

Results in Table 4 show that word-level measures return the same rating independent from the sense being used, thus, they perform good when evaluated on a word-level, but perform poorly on a sense-level. For the word pair *Jaguar—Zoo*, there exist two sense pairs *Atari Jaguar—Zoo* and *Jaguar (animal)—Zoo*. Word-level measures return the same similarity, thus leading to a very low correlation. This was expected, as only sense-based similarity measures can discriminate between different senses of the same word. Somewhat surprisingly, sense-level measures perform also well on a word-level, but their performance increases strongly on sense-level. Our novel measure ESA on senses provides the best results. This is expected as the ambiguous dataset contains many infrequently used senses, which annotators are not aware of.

Our analysis shows that the algorithm for comparing two vectors (i.e. cosine and language model) only influences results for ESA on senses when computed on a word-level. Correlation for Wikipedia Link Measure (WLM) differs depending on whether the overlap of incoming or outgoing links are computed. WLM on word-level using incoming links performs better, while the difference on sense-level evaluation is only marginal. Results show that an evaluation on the level of words and senses may influence performance of measures strongly.

3.3 Pair-wise Evaluation

In a second experiment, we evaluate how well sense-based measures can decide, which one of

[7] Annotators are near-native speakers of English and have university degrees in cultural anthropology and computer science.

[8] The sense of a word is given in parentheses but annotators have access to Wikipedia to get information about those senses.

[9] We report agreement as Krippendorf α with a quadratic weight function.

[10] Although, there exists sense alignment resources, we did not use any alignment.

				Annotator 1		Annotator 2	
Word 1	Word 2	Sense 1	Sense 2	Words	Senses	Words	Senses
Bass	**Fish**	Bass (Fish)	Fish (Animal)	**1**	**4**	**1**	**4**
		Bass (Instrument)			1		1
		Bass (Sound)			1		1
Bass	**Horn**	Bass (Fish)	Horn (Instrument)	**2**	1	**1**	1
		Bass (Instrument)			**3**		**4**
		Bass (Sound)			**3**		3

Table 3: Examples of ratings for two word pairs and all sense combinations with the highest ratings marked bold

		Word-level		Sense-level	
	measure	Spearman	Pearson	Spearman	Pearson
Word measures	ESA	**.456**	.239	-.001	.017
	Lin (WordNet)	.298	.275	.038	.016
Sense measures	ESA on senses (Cosine)	.292	.272	**.642**	.348
	ESA on senses (Lang. Mod.)	.185	.256	**.642**	**.482**
	WLM (out)	.190	.193	.537	.372
	WLM (in)	.287	**.279**	.535	.395

Table 4: Correlation of similarity measures with a human gold standard of ambiguous word pairs.

two sense pairs for one word pair have a higher similarity. We thus create for every word pair all possible sense pairs[11] and count cases where one measure correctly decides, which is the sense pair with a higher similarity.

Table 5 shows evaluation results based on a minimal difference between two sense pairs. We removed all sense pairs with a lower difference of their gold similarity. Column *#pairs* gives the number of remaining sense pairs. If a measure classifies two sense pairs wrongly, it may either be because it rated the sense pairs with an equal similarity or because it reversed the order.

Results show that accuracy increases with increasing minimum difference between sense pairs. Figure 3 emphasizes this finding. Overall, accuracy for this task is high (between .70 and .83), which shows that all the measures can discriminate sense pairs. WLM (out) performs best for most cases with a difference in accuracy of up to .06.

When comparing these results to results from Table 4, we see that correlation does not imply accurate discrimination of sense pairs. Although, ESA on senses has the highest correlation to human ratings, it is outperformed by WLM (out) on the task of discriminating two sense pairs. We see that results are not stable across both evaluation scenarios, however, ESA on senses achieves the highest correlation and performs similar to WLM (out) when comparing sense pairs pair-wise.

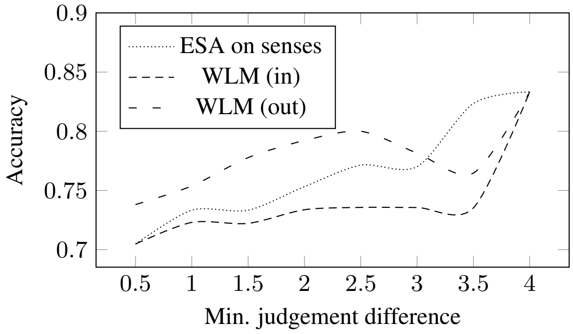


Figure 3: Accuracy distribution depending on minimum difference of similarity ratings

4 Annotation Study II: Re-rating of RG65

We performed a second evaluation study where we asked three human annotators[12] to rate the similarity of word-level pairs in the dataset by Rubenstein and Goodenough (1965). We hypothesize that measures working on the sense-level should have a disadvantage on word-level annotated datasets due to the effects described above that influence annotators towards frequent senses. In our annotation

[11] For one word pair with two senses for one word, there are two possible sense pairs. Three senses result in three sense pairs.

[12] As before, all three annotators are near-native speakers of English and have a university degree in physics, engineering, and computer science.

Min.				Wrong		
diff.	#pairs	measure	Correct	Reverse	Values equal	Accuracy
0.5	420	ESA on senses	296	44	80	.70
		WLM (in)	296	62	62	.70
		WLM (out)	310	76	34	**.74**
1.0	390	ESA on senses	286	38	66	.73
		WLM (in)	282	52	56	.72
		WLM (out)	294	64	32	**.75**
1.5	360	ESA on senses	264	34	62	.73
		WLM (in)	260	48	52	.72
		WLM (out)	280	54	26	**.78**
2.0	308	ESA on senses	232	28	48	.75
		WLM (in)	226	36	46	.73
		WLM (out)	244	46	18	**.79**
2.5	280	ESA on senses	216	22	42	.77
		WLM (in)	206	32	42	.74
		WLM (out)	224	38	18	**.80**
3.0	174	ESA on senses	134	10	30	.77
		WLM (in)	128	20	26	.74
		WLM (out)	136	22	16	**.78**
3.50	68	ESA on senses	56	4	8	**.82**
		WLM (in)	50	6	12	.74
		WLM (out)	52	6	10	.76
4.0	12	ESA on senses	10	2	0	**.83**
		WLM (in)	10	2	0	**.83**
		WLM (out)	10	2	0	**.83**

Table 5: Pair-wise comparison of measures: Results for ESA on senses (language model) and ESA on senses (cosine) do not differ

studies, our aim is to minimize the effect of sense weights.

In previous annotation studies, human annotators could take sense weights into account when judging the similarity of word pairs. Additionally, some senses might not be known by annotators and, thus receive a lower rating. We minimize these effects by asking annotators to select the best sense for a word based on a short summary of the corresponding sense. To mimic this process, we created an annotation tool (see Figure 4), for which an annotator first selects senses for both words, which have the highest similarity. Then the annotator ranks the similarity of these sense pairs based on the complete sense definition.

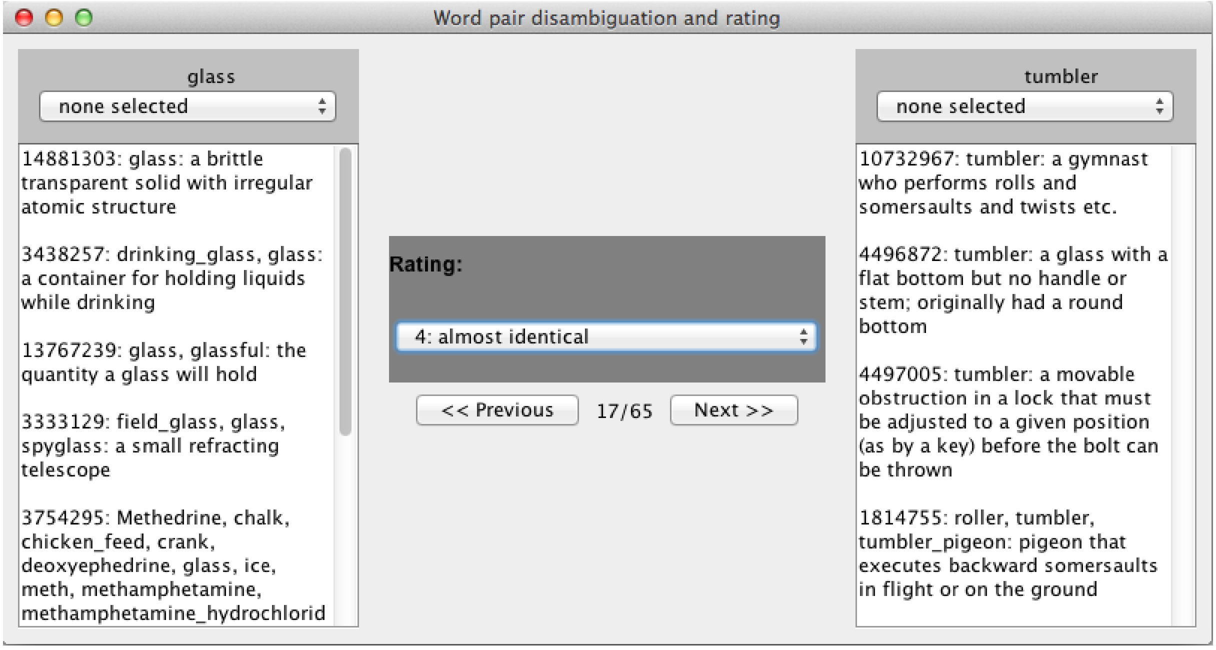
A single word without any context cannot be disambiguated properly. However, when word pairs are given, annotators first select senses based on the second word, e.g. if the word pair is *Jaguar* and *Zoo*, an annotator will select the wild animal for *Jaguar*. After disambiguating, an annotator assigns a similarity score based on both selected senses. To facilitate this process, a definition of each possible sense is shown.

As in the previous experiment, similarity is annotated on a five-point-scale from 0 to 4. Although, we ask annotators to select senses for word pairs, we retrieve only one similarity rating for each word pair, which is the sense combination with the highest similarity.

No sense inventory To compare our results with the original dataset from Rubenstein and Goodenough (1965), we asked annotators to rate similarity of word pairs without any given sense repository, i.e. comparing words directly. The annotators reached an agreement of .73. The resulting gold standard has a high correlation with the original dataset (.923 Spearman and .938 Pearson). This is in line with our expectations and previous work that similarity ratings are stable across time (Bär et al., 2011).

Wikipedia sense inventory We now use the full functionality of our annotation tool and ask annotators to first, select senses for each word and second, rate the similarity. Possible senses and definitions for these senses are extracted from Wikipedia.[13] The same three annotators reached

[13] We use the English Wikipedia version from June 15th, 2010.



Word pair disambiguation and rating

glass

none selected

14881303: glass: a brittle transparent solid with irregular atomic structure

3438257: drinking_glass, glass: a container for holding liquids while drinking

13767239: glass, glassful: the quantity a glass will hold

3333129: field_glass, glass, spyglass: a small refracting telescope

3754295: Methedrine, chalk, chicken_feed, crank, deoxyephedrine, glass, ice, meth, methamphetamine, methamphetamine_hydrochlorid

Rating:

4: almost identical

<< Previous 17/65 Next >>

tumbler

none selected

10732967: tumbler: a gymnast who performs rolls and somersaults and twists etc.

4496872: tumbler: a glass with a flat bottom but no handle or stem; originally had a round bottom

4497005: tumbler: a movable obstruction in a lock that must be adjusted to a given position (as by a key) before the bolt can be thrown

1814755: roller, tumbler, tumbler_pigeon: pigeon that executes backward somersaults in flight or on the ground

Figure 4: User interface for annotation studies: The example shows the word pair *glass—tumbler* with no senses selected. The interface shows WordNet definitons of possible senses in the text field below the sense selection. The highest similarity is selected as sense 4496872 for tumbler is a drinking glass.

an agreement of .66. The correlation to the original dataset is lower than for the re-rating (.881 Spearman, .896 Pearson). This effect is due to many entities in Wikipedia, which annotators would typically not know. Two annotators rated the word pair *graveyard—madhouse* with a rather high similarity because both are names of music bands (still no very high similarity because one is a rock and the other a jazz band).

WordNet sense inventory Similar to the previous experiment, we list possible senses for each word from a sense inventory. In this experiment, we use WordNet senses, thus, not using any named entity. The annotators reached an agreement of .73 and the resulting gold standard has a high correlation with the original dataset (.917 Spearman and .928 Pearson).

Figure 5 shows average annotator ratings in comparison to similarity judgments in the original dataset. All re-rating studies follow the general tendency of having higher annotator judgments for similar pairs. However, there is a strong fluctuation in the mid-similarity area (1 to 3). This is due to fewer word pairs with such a similarity.

4.1 Results & Discussion

We evaluate the similarity measures using Spearman and Pearson correlation with human similarity judgments. We calculate correlations to four human judgments: (i) from the original dataset (Orig.), (ii) from our re-rating study (Rerat.), (iii) from our study with senses from Wikipedia (WP), and (iv) with senses from WordNet (WN). Table 6 shows results for all described similarity measures.

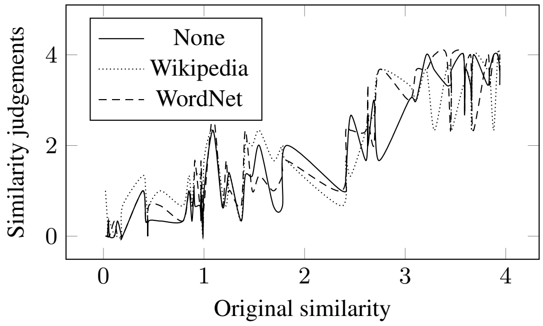


Figure 5: Correlation curve of rerating studies

ESA[14] achieves a Spearman correlation of .751 and a slightly higher correlation (.765) on our re-rating gold standard. Correlation then drops when compared to gold standards with senses from Wikipedia and WordNet. This is expected as the gold standard becomes more sense-aware.

Lin is based on senses in WordNet but still out-

[14] ESA is used with normalized text frequencies, a constant document frequency, and a cosine comparison of vectors.

	Spearman				Pearson			
measure	Orig.	Rerat.	WP	WN	Orig.	Rerat.	WP	WN
ESA	.751	.765	.704	.705	.647	.694	.678	.625
Lin	**.815**	.768	.705	.775	**.873**	**.840**	**.798**	**.846**
ESA on senses (lang. mod.)	.733	.765	.782	.751	.703	.739	.739	.695
ESA on senses (cosine)	.775	**.810**	**.826**	**.795**	.694	.712	.736	.699
WLM (in)	.716	.745	.754	.733	.708	.712	.740	.707
WLM (out)	.583	.607	.652	.599	.548	.583	.613	.568

Table 6: Correlation of similarity measures with a human gold standard on the word pairs by Rubenstein and Goodenough (1965). Best results for each gold standard are marked bold.

performs all other measures on the original gold standard. Correlation reaches a high value for the gold standard based on WordNet, as the same sense inventory for human annotations and measure is applied. Values for Pearson correlation emphasizes this effect: Lin reaches the maximum of .846 on the WordNet-based gold standard.

Correspondingly, the similarity measures ESA on senses and WLM reach their maximum on the Wikipedia-based gold standard. As for the ambiguous dataset in Section 3 ESA on senses outperforms both WLM variants. Cosine vector comparison again outperforms the language model variant for Spearman correlation but impairs it in terms of Pearson correlation. As before WLM (in) outperforms WLM (out) across all datasets and both correlation metrics.

Is word similarity sense-dependent? In general, sense-level similarity measures improve when evaluated with a sense-aware gold standard, while correlation with word-level similarity measures decreases. A further manual analysis shows that sense-level measures perform good when rating very similar word pairs. This is very useful for applications such as information retrieval where a user is only interested in very similar documents.

Our evaluation thus shows that word similarity should not be considered without considering the effect of the used sense inventory. The same annotators rate word pairs differently if they can specify senses explicitly (as seen in Table 3). Correspondingly, results for similarity measures depend on which senses can be selected. Wikipedia contains many entities, e.g. music bands or actors, while WordNet contains fine-grained senses for things (e.g. narrow senses of glass as shown in Figure 4). Using the same sense inventory as the one, which has been used in the annotation process, leads to a higher correlation.

5 Related Work

The work by Schwartz and Gomez (2011) is the closest to our approach in terms of sense annotated datasets. They compare several sense-level similarity measures based on the WordNet taxonomy on sense-annotated datasets. For their experiments, annotators were asked to select senses for every word pair in three similarity datasets. Annotators were not asked to re-rate the similarity of the word pairs, or the sense pairs, respectively. Instead, similarity judgments from the original datasets are used. Possible senses are given by WordNet and the authors report an inter-annotator agreement of .93 for the RG dataset.

The authors then compare Spearman correlation between human judgments and judgments from WordNet-based similarity measures. They focus on differences between similarity measures using the sense annotations and the maximum value for all possible senses. The authors do not report improvements across all measures and datasets. Of ten measures and three datasets, using sense annotations, improved results in nine cases. In 16 cases, results are higher when using the maximum similarity across all possible senses. In five cases, both measures yielded an equal correlation. The authors do not report any overall tendency of results. However, these experiments show that switching from words to senses has an effect on the performance of similarity measures.

The work by Hassan and Mihalcea (2011) is the closest to our approach in terms of similarity measures. They introduce Salient Semantic Analysis (SAS), which is a sense-level measure based on links and disambiguated senses in Wikipedia articles. They create a word-sense-matrix and compute similarity with a modified cosine metric. However, they apply additional normalization factors to optimize for the evaluation metrics which makes a direct comparison of word-level and sense-level variants difficult.

Meyer and Gurevych (2012a) analyze verb similarity with a corpus from Yang and Powers (2006) based on the work by Zesch et al. (2008). They apply variations of the similarity measure ESA by Gabrilovich and Markovitch (2007) using Wikipedia, Wiktionary, and WordNet. Meyer and Gurevych (2012a) report improvements using a disambiguated version of Wiktionary. Links in Wiktionary articles are disambiguated and thus transform the resource to a sense-based resource. In contrast to our work, they focus on the similarity of verbs (in comparison to nouns in this paper) and it applies disambiguation to improve the underlying resource, while we switch the level, which is processed by the measure to senses.

Shirakawa et al. (2013) apply ESA for computation of similarities between short texts. Texts are extended with Wikipedia articles, which is one step to a disambiguation of the input text. They report an improvement of the sense-extended ESA approach over the original version of ESA. In contrast to our work, the text itself is not changed and similarity is computed on the level of texts.

6 Summary and Future Work

In this work, we investigated word-level and sense-level similarity measures and investigated their strengths and shortcomings. We evaluated how correlations of similarity measures with a gold standard depend on the sense inventory used by the annotators.

We compared the similarity measures ESA (corpus-based), Lin (WordNet), and Wikipedia Link Measure (Wikipedia), and a sense-enabled version of ESA and evaluated them with a dataset containing ambiguous terms. Word-level measures were not able to differentiate between different senses of one word, while sense-level measures could even increase correlation when shifting to sense similarities. Sense-level measures obtained accuracies between .70 and .83 when deciding which of two sense pairs has a higher similarity.

We performed re-rating studies with three annotators based on the dataset by Rubenstein and Goodenough (1965). Annotators were asked to first annotate senses from Wikipedia and WordNet for word pairs and then judge their similarity based on the selected senses. We evaluated with these new human gold standards and found that correlation heavily depends on the resource used by the similarity measure and sense repository a human annotator selected. Sense-level similarity measures improve when evaluated with a sense-aware gold standard, while correlation with word-level similarity measures decreases. Using the same sense inventory as the one, which has been used in the annotation process, leads to a higher correlation. This has implications for creating word similarity datasets and evaluating similarity measures using different sense inventories.

In future work we would like to analyze how we can improve sense-level similarity measures by disambiguating a large document collection and thus retrieving more accurate frequency values. This might reduce the sparsity of term-document-matrices for ESA on senses. We plan to use word sense disambiguation components as a preprocessing step to evaluate whether sense similarity measures improve results for text similarity. Additionally, we plan to use sense alignments between WordNet and Wikipedia to enrich the term-document matrix with additional links based on semantic relations.

The datasets, annotation guidelines, and our experimental framework are publicly available in order to foster future research for computing sense similarity.[15]

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, by the Klaus Tschira Foundation under project No. 00.133.2008, and by the German Federal Ministry of Education and Research (BMBF) within the context of the Software Campus project *open window* under grant No. 01IS12054. The authors assume responsibility for the content. We thank Pedro Santos, Michèle Spankus and Markus Bücker for their valuable contribution. We thank the anonymous reviewers for their helpful comments.

[15] `www.ukp.tu-darmstadt.de/data/text-similarity/sense-similarity/`

References

Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet. In *Computational Linguistics and Intelligent Text*, pages 136—-145.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A Reflective View on Text Similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 515–520, Hissar, Bulgaria.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611.

Gregory Grefenstette. 1992. Sextant: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 324—-326, Newark, Delaware, USA. Association for Computational Linguistics.

Samer Hassan and Rada Mihalcea. 2011. Semantic Relatedness Using Salient Semantic Analysis. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence, (AAAI 2011)*, pages 884–889, San Francisco, CA, USA.

Jay J Jiang and David W Conrath. 1997. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of 10th International Conference Research on Computational Linguistics*, pages 1–15.

Dekang Lin. 1998. An Information-theoretic Definition of Similarity. In *In Proceedings of the International Conference on Machine Learning*, volume 98, pages 296—-304.

Christian M. Meyer and Iryna Gurevych. 2012a. To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1763–1780, Mumbai, India.

Christian M. Meyer and Iryna Gurevych. 2012b. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford University Press, Oxford, UK, November.

David Milne and Ian H Witten. 2008. Learning to Link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 509—-518.

David Milne. 2007. Computing Semantic Relatedness using Wikipedia Link Structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Philip Resnik. 1999. Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130.

Herbert Rubenstein and John B Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627—-633.

Hansen A Schwartz and Fernando Gomez. 2011. Evaluating Semantic Metrics on Tasks of Concept Similarity. In *FLAIRS Conference*.

Nuno Seco, Tony Veale, and Jer Hayes. 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proceedings of European Conference for Artificial Intelligence*, number Ic, pages 1089–1093.

Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2013. Probabilistic Semantic Similarity Measurements for Noisy Short Texts using Wikipedia Entities. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 903–908, New York, New York, USA. ACM Press.

Dongqiang Yang and David MW Powers. 2006. Verb Similarity on the Taxonomy of WordNet. In *Proceedings of GWC-06*, pages 121—-128.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 861–867, Chicago, IL, USA.