Performance Evaluation of WordNet-based Semantic Relatedness Measures for Word Prediction in Conversational Speech

Michael Pucher, FTW, pucher@ftw.at

December 2, 2004

1 Abstract

The performance of eight WordNet-based semantic similarity/relatedness measures for word prediction in conversational speech was evaluated. We give a ranking of the different measures which shows that the performance of the measures differs significantly for noun and verb prediction. We also varied the dialog context and used cross part-of-speech comparison.

2 Introduction

The recognition of conversational speech is a hard problem. Semantic relatedness measures can improve speech recognition performance when using contextual information, as Demetriou [5] has shown. The standard n-gram approach in language modeling for speech recognition cannot cope with long distance dependencies [4]. Therefore J. Bellegarda [2] proposed combining n-gram language models, which are effective for predicting local dependencies, with latent semantic analysis for long distance dependencies. WordNet-based semantic relatedness measures can be used for word prediction using long distance dependencies, as in these examples from our experiments:

(1) B: I I well, you should see what the ⌊students⌋
B: after they torture them for six ⌊years⌋ in middle ⌊school⌋ and high ⌊school⌋ they don't want to do anything in ⌊**college**⌋ particular.

In this example the word *college* can be predicted from the noun context using semantic relatedness measures, here between *students* and *college*. A 3-gram model would give a ranking for *college* in the context of *anything in*. An 8-gram would predict *college* from *they don't want to do anything in*, but the strongest predictor is *students*.

(2) B: everyone who's who's extra busy, of course, you know who's ⌊doing⌋ the ⌊cooking⌋, like tonight it was Benny and me.

A: mm.

B: I ⌊mean⌋ e- so all the ⌊people⌋ who are ⌊**working**⌋.

In example (2) *working* can be predicted from *people*, *cooking* and *doing*, since for verbs we use the verbs and nouns in the context.

In addition to such predictions based on semantic relatedness there is another type of prediction which relies on WordNet's morphological analyzer. In these predictions a word is predicted if the word itself or an inflection occurs in the context.

Many different relatedness/similarity measures for WordNet have been proposed. Here we evaluate the performance of these measures for word prediction in conversational speech. We want to use these measures for speech recognition hypothesis rescoring, which can be done on word graphs or n-best lists. For the rescoring of word graphs one has to proceed in a left-to-right manner, while it is possible to use the whole sentence as a context when rescoring n-best lists. Therefore we defined two context measures for the performance evaluation. The first measure defines the relatedness of a word and a context (Definition 4) and can be used for word graph and n-best list rescoring. The second measure defines the relatedness between a word, a sentence, and a context (Definition 6) and can be used for rescoring of n-best lists, where the whole sentence can be used as an additional context for measuring the relatedness. For the sake of simplicity we use the term sentence here. Actually we add the bag of words in a dialog turn in the CallHome corpus to the context, which is not necessarily a sentence in any syntactic sense.

As an evaluation corpus we used five dialogs from the English CallHome corpus. The corpus was tagged using a trigram tagger and the Brown Corpus, and the content words (nouns, verbs) were extracted.

Most relatedness measures do not work across different parts of speech, e.g. one cannot calculate the relatedness between a verb and a noun directly, so we did not use cross part-of-speech comparison in the first run. WordNet has four part-of-speech tags - nouns, verbs, adjectives and adverbs. This is already simplified in comparison to the output of the part-of-speech tagger. We used only nouns and verbs because most measures do not work for adjectives and adverbs.

We want to apply our results to multiparty dialogs. Therefore we used the whole dialog context (two speakers), as well as the subdialog contexts (one speaker), which are only the monologs in our case.

3 WordNet-Based Semantic Relatedness Measures

For our evaluation we used eight relatedness/similarity measures from the *Perl* package *WordNet Similarity* written by T. Pedersen et.al. [12]. The measures `res` [13], `lin` [10] and `jcn` [7] are based on the information content, the measures `lch` [9], `wup` [14] and `path` use path lengths between two words in the WordNet graph, and `hso` [6] and `lesk` [1] allow for comparison across part-of-speech boundaries.

Here we will only explain the relatedness measures that perform best in our experiments. First we define the *information content* of a concept and the *least common subsumer* (LCS), where:

> *...the LCS of concepts A and B is the most specific concept that is an ancestor of both A and B.* [12]

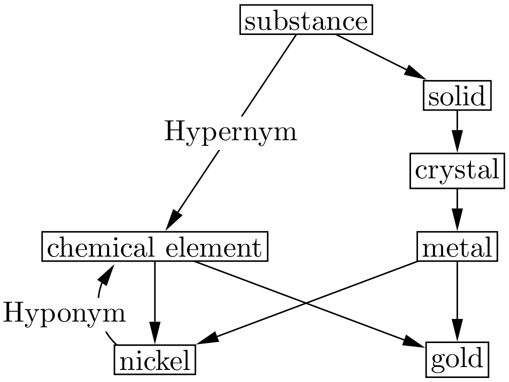

Figure 1: Fragment of WordNet taxonomy

In Figure 1 common subsumers of the concepts *nickel* and *gold* are *chemical element*, *metal*, *substance* etc. But because WordNet allows multiple inheritance *nickel* and *gold* have two least common subsumers *chemical element* and *metal*.

Definition 1 *Information content*

$$\mathrm{IC}(c) \triangleq -\log(\frac{\mathrm{freq}(c)}{N})$$

The information content of a concept IC(c) is defined as the negative log likelihood of the probability of encountering an instance of the concept. The probability of a concept c is given by the frequency of c in the corpus freq(c) divided by the number of concepts in the corpus N. The more specific a concept, the higher its information content.

Because WordNet allows multiple inheritence `res` takes the least common subsumer with the highest information content. IC(c_i) is the *information content* of c_i and LCS(c_i, c_j) are the LCS of c_i and c_j.

Definition 2 *Resnik measure (`res`)*

$$\mathrm{rel}_{\mathtt{res}}(c_1, c_2) \triangleq \max_{c \in \mathrm{LCS}(c_1, c_2)}(\mathrm{IC}(c))$$

The `jcn` measure additionally uses the information content of the concepts that are compared. The distance between two concepts is defined as:

Definition 3 *Jiang and Conrath measure (`jcn`)*

$$\mathrm{rel}_{\mathtt{jcn}}(c_1, c_2) \triangleq \mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 * \max_{c \in \mathrm{LCS}(c_1, c_2)}(\mathrm{IC}(c))$$

In the *WordNet Similarity* package `jcn` is implemented as a similarity measure.

The `lesk` (Banerjee and Pedersen) measure uses the number of identical words in the extended WordNet glosses of two words as a measure of relatedness.

The `path` measure uses the shortest path between two words in the WordNet graph. We will use the term *relatedness measure* because it is more general and subsumes distance as well as similarity measures.

4 Word Context Relatedness

Since we want to measure the semantic relatedness of a word and a context we have to define a word-context relatedness measure that uses the WordNet measures. For the first evaluation we used a slightly modified version of

Kozima and Ito's definition [8] of semantic relatedness. They used a semantic vector space, so they could directly define the distance between two words in context $\operatorname{dist}(w, w' \mid C)$ by taking the vector distance. A context C is a multiset consisting of the previous δ words in the dialog, where δ is the context width.

In our case $\operatorname{rel}(w, w')$ is one of the WordNet based relatedness measures. For the first evaluation we did not use cross part-of-speech measuring, the Brown corpus was the basis for the information content files needed for the `lin`, `res` and `jcn` measures, and we took the whole dialog (e.g. both speakers in each dialog of the CallHome corpus) as the context.

The relatedness of a word and a context is defined as the sum of the relatedness of the word and all words in the context.

Definition 4 *Word-context relatedness*

$$\operatorname{rel}_{\mathrm{W}}(w \mid C) = \frac{1}{\mid C \mid} \sum_{w' \in C} \operatorname{rel}(w, w')$$

4.1 Performance Measuring

To calculate the performance of a measure we used the method described in [8]. We used this performance measure because it allows us to compare differently scaled semantic relatedness measures. To measure the word prediction performance for a word w_p in a context C, the vocabulary $V = \{w_1, \ldots, w_n\}$ of the whole dialog is ordered according to each word's relatedness to the context so that

$$\operatorname{rel}_{\mathrm{W}}(w_{i_1} \mid C) > \operatorname{rel}_{\mathrm{W}}(w_{i_2} \mid C) > \cdots > \operatorname{rel}_{\mathrm{W}}(w_{i_n} \mid C)$$

for all n words of the vocabulary V. Suppose $k = i_p$ is the position of word w_p in this ordering. The performance for w_p is:

Definition 5 *Performance Measure*

$$\operatorname{perf}(w_p) = \frac{\mid V \mid /2 - k}{\mid V \mid /2}$$

If the word w_p occurs in the first half of the ordered vocabulary list, the performance score is positive. If it occurs in the second half it is negative. The performance scores are between -1 and 1. Using this measure we calculated the following scores for the eight different WordNet based relatedness measures. For distance measures the ordering has to be reversed.

4.2 Evaluation Results

We used a context-width $\delta = 5$ in all our performance measures. We used five dialogs from the English CallHome corpus, and calculated the average performance value for verbs and nouns.

We also ran tests with $\delta = 10$ and $\delta = 15$; the average scores were slightly higher, but the ranking of the measures did not change. Due to the computationally expensive implementation of some measures, we took the lowest context-width.

The evaluation tables contain the name of the relatedness measure, the part-of-speech (N for noun and V for verb), and the mean performance values for each part of speech. The last column contains the mean performance value for nouns and verbs together and determines the ranking of the measures. As one can see from Table 1 the measure `jcn` has the best overall performance, followed by the measures `wup` and `path` which are based on path lenghts.

Rel	POS	Performance	
`jcn`	N	0.387	0.385
`jcn`	V	0.383	
`wup`	N	0.299	0.313
`wup`	V	0.328	
`path`	N	0.333	0.307
`path`	V	0.281	
`lesk`	N	0.299	0.288
`lesk`	V	0.277	
`res`	N	0.290	0.288
`res`	V	0.286	
`hso`	N	0.254	0.227
`hso`	V	0.201	
`lch`	N	0.250	0.225
`lch`	V	0.200	
`lin`	N	0.220	0.194
`lin`	V	0.169	

Table 1: Word-context relatedness performance

Rel	POS	Performance	
`lesk`	N	-0.002	0.212
`lesk`	V	0.427	
`hso`	N	0.068	0.186
`hso`	V	0.305	

Table 2: Word-context relatedness performance across POS

It is surprising that the `wup` and `path` measure, which are only based on path lenghts have such good performance scores. The good performance of the jcn measure was already reported by Budanitsky [3] for the task of malapropism correction.

Figure 2 shows the mean performance of some measures for the five dialogs, for all dialogs, and for noun and verb measures together. We can see that the performance varies significantly from dialog to dialog. Since the performance score is always bigger than zero, we can conclude that the measures perform better than random.

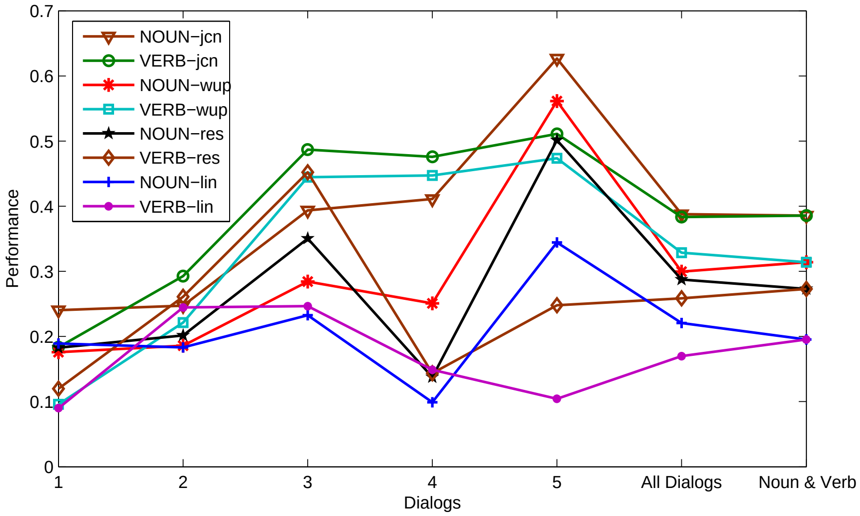

Figure 2: Word-context relatedness performance

5 Crossing Part-of-Speech Boundaries

Table 2 shows the results for the two measures that allow cross part-of-speech comparison, namely hso and lesk, for which the context contains nouns and verbs.

The lesk measure performs very well for verbs and is unusable for nouns, which gives an overall performance score of 0.212. The prediction of verbs from a context containing nouns and verbs performs better than from a context containing only verbs. The prediction of nouns from a mixed context performs worse than from a context containing only nouns, which is shown in the noun column of lesk in Table 2.

The hso measure also performs worse for nouns when using a mixed context (see Table 2) and better for verbs, but it is still outperformed by the lesk measure.

6 Changing Dialog Context

Semantic relatedness measures can also be used for speech recognition of multiparty dialogs. This section evaluates the performance of the measures for different dialog contexts. The CallHome corpus only includes dialogs between two speakers. The words in the context can come either from the whole dialog or just from the monolog.

Therefore we divided a dialog D into a set of monologs $\{M_1 \dots M_n\}$ each monolog corresponding to one speaker. The set of all subdialogs is given by the power set of D, $\mathcal{P}(D)$. The context words can be restricted to each of these subdialogs.

Since we consider dialogs with two speakers only, the possible subdialogs are the monologs, and the whole dialog. For our evaluation we used the word-context relatedness and resctricted the context to the monologs (The performance of word-context relatedness for the whole dialog is already shown in Table 1).

As Table 4 shows, the performance decreases in general when using just the monolog context, relative to Table 1, which uses the whole dialog as a context. Only the `jcn` measure still performs quite good when using just the monolog.

7 Word Sentence Context Relatedness

The performance of the word-context relatedness gives us a hint how well the measures will work for algorithms that work in a left-to-right manner, e.g. in the rescoring of word graphs or n-best lists. For the rescoring of word graphs one has to extend the definition to the relatedness of two words in context, which can then be used to rescore the transition probabilities in the word graph. For n-best lists one has to calculate the relatedness for each sentence in the list.[1]

For the rescoring of n-best lists it is however not necessary to proceed in a left-to-right manner. The word-sentence-context relatedness can be used for the rescoring of n-best lists. This relatedness does not only use the context of the preceeding words, but the whole sentence.

It can be defined in the following way: Suppose we have a sentence $S =< w_1, \dots, w_n >$. Let $\text{pre}(w_i, S)$ be the set $\bigcup_{j<i} w_j$ and $\text{post}(w_i, S)$ be

[1] It is also necessary to translate the non-probabilistic relatedness measures into probabilistic measures.

the set $\bigcup_{j>i} w_j$. Then we can define the word-sentence-context relatedness as:

Definition 6 *Word-sentence-context relatedness*

$$\text{rel}_\text{S}(w, S \mid C) = \text{rel}_\text{W}(w \mid \text{pre}(w, S) \cup \text{post}(w, S) \cup C)$$

Rel	POS	Performance	
`path`	N	0.392	0.398
`path`	V	0.405	
`jcn`	N	0.367	0.367
`jcn`	V	0.368	
`res`	N	0.344	0.356
`res`	V	0.369	
`lesk`	N	0.247	0.295
`lesk`	V	0.343	
`lch`	N	0.265	0.236
`lch`	V	0.207	
`hso`	N	0.248	0.221
`hso`	V	0.195	
`lin`	N	0.220	0.194
`lin`	V	0.169	
`wup`	N	0.206	0.184
`wup`	V	0.163	

Table 3: Word-sentence-context relatedness performance

Rel	POS	Performance	
`jcn`	N	0.334	0.315
`jcn`	V	0.297	
`path`	N	0.256	0.222
`path`	V	0.188	
`lch`	N	0.249	0.217
`lch`	V	0.186	
`lesk`	N	0.237	0.210
`lesk`	V	0.183	
`hso`	N	0.230	0.200
`hso`	V	0.171	
`res`	N	0.214	0.183
`res`	V	0.153	
`lin`	N	0.192	0.167
`lin`	V	0.143	
`wup`	N	0.184	0.164
`wup`	V	0.144	

Table 4: Word-monolog-context relatedness performance

When using this measure the context width is $\delta = 5$ plus the number of verbs/nouns in the sentence the word belongs to. So we expected the performance to be better than with the word-context measure. As Table 3 shows, this is not the case in general. The `path` and `res` measure perform better for nouns and verbs, and the `lesk` measure performs better for verbs.

8 Performance Comparison

8.1 Noun Measures

T-tests for paired samples indicated that the performance values of the `path` measure using the word-sentence-context and the whole dialog (`path_d_s`) were significantly higher ($p < .05, \#\text{nouns} = 722$) than all other noun-related measures , except the jcn measure using the word-context and the whole dialog (jcn_d_w). Revealing the second-highest mean performance, jcn_d_w performed significantly better than most other noun-related measures (paired samples t-tests; $p < .05, \#\text{nouns} = 722$), except path_d_s, res_d_s, jcn using the word-context of the monolog (jcn_m_w) and jcn using the word-sentence-context (jcn_d_s). We can conclude that the jcn measure performs best when using the word-context, but there is no significant difference if the whole dialog or just the monolog is used.

8.2 Verb Measures

According to t-tests for paired samples, lesk using a mixed word-context of the dialog (lesk_cross_d_w) performed significantly better than all other verb-related measures ($p = .05, \#\text{verbs} = 597$), except path_d_s, which had the second-highest mean performance value. It has to be evaluated if the lesk measure using a mixed context and the word-sentence-context (lesk_cross_d_s) outperforms path_d_s.

9 Conclusion

In this paper we have given a ranking of the usefulness of different semantic similarity measures based on WordNet for word prediction in conversational speech. We have shown that there are significant differences in the performance of these measures.

All measures perform better than random on this task, with the jcn measure performing best for nouns using the word-context of the whole dialog or the monolog, and the path measure performing best for nouns using the word-sentence-context of the dialog. The same result for the jcn measure was obtained by Budanitsky [3] for a different task. The lesk measure performs best for verbs using a mixed word-context. We can conclude that different measures should be used for the prediction of nouns and verbs, and for different contexts.

These results should allow an investigation of the use of the best performing measures for the task of speech recognition hypotheses rescoring for multiparty dialogs.

10 Acknowledgements

The work presented here is part of the project MONA [11] carried out at the Telecommunications Research Center Viennna (ftw.) and is supported by Kapsch Carrier-Com AG, Mobilkom Austria AG and Siemens Österreich AG together with the Austrian competence centre programme **K***plus*.

References

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th Int. Joint Conf. on Artificial Intelligence*, pages 805–810, Acapulco, August 2003.

[2] J. Bellegarda. Large vocabulary speech recognition with multispan statistical language models. In *IEEE Transactions on Speech and Audio Processing, Vol 8.*, January 2000.

[3] A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, 2nd of the North American Chapter of the ACL*, Pittsburgh, 2001.

[4] C. Chelba and F. Jelinek. Exploiting syntactic structure for language modeling. In *Proceedings COLING-ACL'98*, Montreal, Canada, 1998.

[5] G. Demetriou, E. Atwell, and C. Souter. Knowledge from machine readable dictionaries for domain independent language modelling. In *Proc. of LREC 2000, 2nd International Conference on Language Resources and Evaluation*, 2000.

[6] G. Hirst and D. St-Onge. Lexical chains as representations of context for the deduction and correction of malapropisms. In *WordNet:An electronical lexical database*, pages 305–332. MIT Press, 1998.

[7] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan, 1997.

[8] H. Kozima and A. Ito. Context-sensitive measurement of word distance by adaptive scaling of a semantic space. In *Proc. of the International Conference "Recent Advances in Natural Language Processing", RANLP-95*, pages 161–168, Bulgaria, 1995.

[9] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In *WordNet: An electronic lexical database*, pages 265–283. MIT Press, 1998.

[10] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, Madison, August 1998.

[11] Mona - Mobile Multimodal Next Generation Applications, http://mona.ftw.at.

[12] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *Proc. of Fifth Annual Meeting of the North American Chapter of the ACL (NAACL-04)*, Boston, MA, 2004.

[13] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

[14] Z. Wu and M. Palmer. Verb semantics and lexical selection. pages 133–138, Las Cruces, Mexico, 1994.