Investigating Lexical Substitution Scoring for Subtitle Generation

Oren Glickman and **Ido Dagan**
Computer Science Department
Bar Ilan University
Ramat Gan, Israel
{glikmao,dagan}@cs.biu.ac.il

Mikaela Keller and **Samy Bengio**
IDIAP Research Institute
Martigny,
Switzerland
{mkeller,bengio}@idiap.ch

Walter Daelemans
CNTS
Antwerp, Belgium
walter.daelemans@ua.ac.be

Abstract

This paper investigates an isolated setting of the lexical substitution task of replacing words with their synonyms. In particular, we examine this problem in the setting of subtitle generation and evaluate state of the art scoring methods that predict the validity of a given substitution. The paper evaluates two context independent models and two contextual models. The major findings suggest that distributional similarity provides a useful complementary estimate for the likelihood that two Wordnet synonyms are indeed substitutable, while proper modeling of contextual constraints is still a challenging task for future research.

1 Introduction

Lexical substitution - the task of replacing a word with another one that conveys the same meaning - is a prominent task in many Natural Language Processing (NLP) applications. For example, in query expansion for information retrieval a query is augmented with synonyms of the original query words, aiming to retrieve documents that contain these synonyms (Voorhees, 1994). Similarly, lexical substitutions are applied in question answering to identify answer passages that express the sought answer in different terms than the original question. In natural language generation it is common to seek lexical alternatives for the same meaning in order to reduce lexical repetitions. In general, lexical substitution aims to preserve a desired meaning while coping with the lexical variability of expressing that meaning. Lexical substitution can thus be viewed within the general framework of recognizing entailment between text segments (Dagan et al., 2005), as modeling entailment relations at the lexical level.

In this paper we examine the lexical substitution problem within a specific setting of text compression for subtitle generation (Daelemans et al., 2004). Subtitle generation is the task of generating target language TV subtitles for video recordings of a source language speech. The subtitles should be of restricted length, which is often shorter than the full translation of the original speech, yet they should maintain as much as possible the meaning of the original content. In a typical (automated) subtitling process the original speech is first translated fully into the target language and then the target translation is compressed to optimize the length requirements. One of the techniques employed in the text compression phase is to replace a target language word in the original translation with a shorter synonym of it, thus reducing the character length of the subtitle. This is a typical lexical substitution task, which resembles similar operations in other text compression and generation tasks (e.g. (Knight and Marcu, 2002)).

This paper investigates the task of assigning likelihood scores for the correctness of such lexical substitutions, in which words in the original translation are replaced with shorter synonyms. In our experiments we use WordNet as a source of candidate synonyms for substitution. The goal is to score the likelihood that the substitution is admissible, i.e. yielding a valid sentence that preserves the original meaning. The focus of this paper is thus to utilize the subtitling setting in order to investigate lexical substitution models in isolation, unlike most previous literature in which this sub-task has been embedded in larger systems and was not evaluated directly.

We examine four statistical scoring models, of two types. Context independent models score the general likelihood that the original word is "replaceable" with the candidate synonym, in an arbitrary context. That is, trying to filter relatively bizarre synonyms, often of rare senses, which are abundant in WordNet but are unlikely to yield valid substitutions. Contextual models score the "fitness" of the replacing word within the context of the sentence, in order to filter out synonyms of senses of the original word that are not the right sense in the given context.

We set up an experiment using actual subtitling data and human judgements and evaluate the different scoring methods. Our findings suggest the dominance, in this setting, of generic context-independent scoring. In particular, considering distributional similarity amongst WordNet synonyms seems effective for identifying candidate substitutions that are indeed likely to be applicable in actual texts. Thus, while distributional similarity alone is known to be too noisy as a sole basis for meaning-preserving substitutions, its combination with WordNet allows reducing the noise caused by the many WordNet synonyms that are unlikely to correspond to valid substitutions.

2 Background and Setting

2.1 Subtitling

Automatic generation of subtitles is a summarization task at the level of individual sentences or occasionally of a few contiguous sentences. Limitations on reading speed of viewers and on the size of the screen that can be filled with text without the image becoming too cluttered, are the constraints that dynamically determine the amount of compression in characters that should be achieved in transforming the transcript into subtitles. Subtitling is not a trivial task, and is expensive and time-consuming when experts have to carry it out manually. As for other NLP tasks, both statistical (machine learning) and linguistic knowledge-based techniques have been considered for this problem. Examples of the former are (Knight and Marcu, 2002; Hori et al., 2002), and of the latter are (Grefenstette, 1998; Jing and McKeown, 1999). A comparison of both approaches in the context of a Dutch subtitling system is provided in (Daelemans et al., 2004). The required sentence simplification is achieved either by deleting material, or by paraphrasing parts of the sentence into shorter expressions with the same meaning. As a special case of the latter, lexical substitution is often used to achieve a compression target by substituting a word by a shorter synonym. It is on this subtask that we focus in this paper. Table 1 provides a few examples. E.g. by substituting "happen" by "occur" (example 3), one character is saved without affecting the sentence meaning .

2.2 Experimental Setting

The data used in our experiments was collected in the context of the MUSA (Multilingual Subtitling of Multimedia Content) project (Piperidis et al., 2004)[1] and was kindly provided for the current study. The data was provided by the BBC in the form of *Horizon* documentary transcripts with the corresponding audio and video. The data for two documentaries was used to create a dataset consisting of sentences from the transcripts and the corresponding substitution examples in which selected words are substituted by a shorter Wordnet synonym. More concretely, a *substitution example* thus consists of an original sentence $s = w_1 \ldots w_i \ldots w_n$, a specific *source* word w_i in the sentence and a *target* (shorter) WordNet synonym w' to substitute the source. See Table 1 for examples. The dataset consists of 918 substitution examples originating from 231 different sentences.

An annotation environment was developed to allow efficient annotation of the substitution examples with the classes *true* (admissible substitution, in the given context) or *false* (inadmissible substitution). About 40% of the examples were judged as true. Part of the data was annotated by an additional annotator to compute annotator agreement. The Kappa score turned out to be 0.65, corresponding to "Substantial Agreement" (Landis and Koch, 1997). Since some of the methods we are comparing need tuning we held out a random subset of 31 original sentences (with 121 corresponding examples) for development and kept for testing the resulting 797 substitution ex-

[1] http://sinfos.ilsp.gr/musa/

id	sentence	source	target	judgment
1	The answer may be found in the behaviour of animals.	answer	reply	false
2	... and the answer to that was - Yes	answer	reply	true
3 4 5	We then wanted to know what would happen if we delay the movement of the subject's left hand	happen subject subject	occur topic theme	true false false
6	people weren't laughing they were going stone sober.	stone	rock	false
7	if we can identify a place where the seizures are coming from then we can go in and remove just that small area.	identify	place	false
8	my approach has been the first to look at the actual structure of the laugh sound.	approach	attack	false
9	He quickly ran into an unexpected problem.	problem	job	false
10	today American children consume 5 times more Ritalin than the rest of the world combined	consume	devour	false

Table 1: Substitution examples from the dataset along with their annotations

amples from the remaining 200 sentences.

3 Compared Scoring Models

We compare methods for scoring lexical substitutions. These methods assign a score which is expected to correspond to the likelihood that the synonym substitution results in a valid subtitle which preserves the main meaning of the original sentence.

We examine four statistical scoring models, of two types. The context independent models score the general likelihood that the source word can be replaced with the target synonym regardless of the context in which the word appears. Contextual models, on the other hand, score the fitness of the target word within the given context.

3.1 Context Independent Models

Even though synonyms are substitutable in theory, in practice there are many rare synonyms for which the likelihood of substitution is very low and will be substitutable only in obscure contexts. For example, although there are contexts in which the word *job* is a synonym of the word *problem*[2], this is not typically the case and overall job is not a good target substitution for the source *problem* (see example 9 in Table 1). For this reason synonym thesauruses such as WordNet tend to be rather noisy for practical purposes, raising the need to score such synonym substitutions and accordingly prioritize substitutions that are more likely to be valid in an arbitrary context.

As representative approaches for addressing this problem, we chose two methods that rely on statistical information of two types: supervised sense distributions from SemCor and unsupervised distributional similarity.

3.1.1 WordNet based Sense Frequencies (semcor)

The obvious reason that a target synonym cannot substitute a source in some context is if the source appears in a different sense than the one in which it is synonymous with the target. This means that a priori, synonyms of frequent senses of a source word are more likely to provide correct substitutions than synonyms of the word's infrequent senses.

To estimate such likelihood, our first measure is based on sense frequencies from SemCor (Miller et al., 1993), a corpus annotated with Wordnet senses. For a given source word u and target synonym v the score is calculated as the percentage of occurrences of u in SemCor for which the annotated synset contains v (i.e. u's occurrences in which its sense is synonymous with v). This corresponds to the prior probability estimate that an occurrence of u (in an arbitrary context) is actually a synonym of v. Therefore it is suitable as a prior score for lexical substitution.[3]

3.1.2 Distributional Similarity (sim)

The SemCor based method relies on a supervised approach and requires a sense annotated corpus. Our

[2]WordNet lists job as a possible member of the synset for a state of difficulty that needs to be resolved, as might be used in sentences like "it is always a job to contact him"

[3]Note that WordNet semantic distance measures such as those compared in (Budanitsky and Hirst, 2001) are not applicable here since they measure similarity between synsets rather than between synonymous words within a single synset.

second method uses an unsupervised distributional similarity measure to score synonym substitutions. Such measures are based on the general idea of Harris' Distributional Hypothesis, suggesting that words that occur within similar contexts are semantically similar (Harris, 1968).

As a representative of this approach we use Lin's dependency-based distributional similarity database. Lin's database was created using the particular distributional similarity measure in (Lin, 1998), applied to a large corpus of news data (64 million words) [4]. Two words obtain a high similarity score if they occur often in the same contexts, as captured by syntactic dependency relations. For example, two verbs will be considered similar if they have large common sets of modifying subjects, objects, adverbs etc.

Distributional similarity does not capture directly meaning equivalence and entailment but rather a looser notion of meaning similarity (Geffet and Dagan, 2005). It is typical that non substitutable words such as antonyms or co-hyponyms obtain high similarity scores. However, in our setting we apply the similarity score only for WordNet synonyms in which it is known a priori that they are substitutable is some contexts. Distributional similarity may thus capture the statistical degree to which the two words are substitutable in practice. In fact, it has been shown that prominence in similarity score corresponds to sense frequency, which was suggested as the basis for an unsupervised method for identifying the most frequent sense of a word (McCarthy et al., 2004).

3.2 Contextual Models

Contextual models score lexical substitutions based on the context of the sentence. Such models try to estimate the likelihood that the target word could potentially occur in the given context of the source word and thus may replace it. More concretely, for a given substitution example consisting of an original sentence $s = w_1 \dots w_i \dots w_n$, and a designated source word w_i, the contextual models we consider assign a score to the substitution based solely on the target synonym v and the context of the source word in the original sentence, $\{w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n\}$, which is represented in a bag-of-words format.

Apparently, this setting was not investigated much in the context of lexical substitution in the NLP literature. We chose to evaluate two recently proposed models that address exactly the task at hand: the first model was proposed in the context of lexical modeling of textual entailment, using a generative Naïve Bayes approach; the second model was proposed in the context of machine learning for information retrieval, using a discriminative neural network approach. The two models were trained on the (unannotated) sentences of the BNC 100 million word corpus (Burnard, 1995) in bag-of-words format. The corpus was broken into sentences, tokenized, lemmatized and stop words and tokens appearing only once were removed. While training of these models is done in an unsupervised manner, using unlabeled data, some parameter tuning was performed using the small development set described in Section 2.

3.2.1 Bayesian Model (bayes)

The first contextual model we examine is the one proposed in (Glickman et al., 2005) to model textual entailment at the lexical level. For a given target word this unsupervised model takes a binary text categorization approach. Each vocabulary word is considered a class, and contexts are classified as to whether the given target word is likely to occur in them. Taking a probabilistic Naïve-Bayes approach the model estimates the conditional probability of the target word given the context based on corpus co-occurrence statistics. We adapted and implemented this algorithm and trained the model on the sentences of the BNC corpus.

For a bag-of-words context $C = \{w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n\}$ and target word v the Naïve Bayes probability estimation for the conditional probability of a word v may occur in a given a context C is as follows:

$$\mathrm{P}(v|C) = \frac{\mathrm{P}(C|v)\,\mathrm{P}(v)}{\mathrm{P}(C|v)\,\mathrm{P}(v)+\mathrm{P}(C|\neg v)\,\mathrm{P}(\neg v)} \approx \frac{\mathrm{P}(v)\prod_{w\in C}\mathrm{P}(w|v)}{\mathrm{P}(v)\prod_{w\in C}\mathrm{P}(w|v)+\mathrm{P}(\neg v)\prod_{w\in C}\mathrm{P}(w|\neg v)} \tag{1}$$

where $\mathrm{P}(w|v)$ is the probability that a word w appears in the context of a sentence containing v and correspondingly $\mathrm{P}(w|\neg v)$ is the probability that w

[4] available at http://www.cs.ualberta.ca/~lindek/downloads.htm

appears in a sentence not containing v. The probability estimates were obtained from the processed BNC corpus as follows:

$$\mathrm{P}(w|v) = \frac{|w \text{ appears in sentences containing } v|}{|\text{words in sentences containing } v|}$$
$$\mathrm{P}(w|\neg v) = \frac{|w \text{ occurs in sentences not containing } v|}{|\text{words in sentences not containing } v|}$$

To avoid 0 probabilities these estimates were smoothed by adding a small constant to all counts and normalizing accordingly. The constant value was tuned using the development set to maximize average precision (see Section 4.1). The estimated probability, $\mathrm{P}(v|C)$, was used as the confidence score for each substitution example.

3.2.2 Neural Network Model (nntr)

As a second contextual model we evaluated the Neural Network for Text Representation (NNTR) proposed in (Keller and Bengio, 2005). NNTR is a discriminative approach which aims at modeling how likely a given word v is in the context of a piece of text C, while learning a more compact representation of reduced dimensionality for both v and C.

NNTR is composed of 3 Multilayer Perceptrons, noted $mlp_{\mathcal{A}}()$, $mlp_{\mathcal{B}}()$ and $mlp_{\mathcal{C}}()$, connected as follow:

$$NNTR(v, C) = mlp_{\mathcal{C}}[mlp_{\mathcal{A}}(v), mlp_{\mathcal{B}}(C)].$$

$mlp_{\mathcal{A}}(v)$ and $mlp_{\mathcal{B}}(C)$ project respectively the vector space representation of the word and text into a more compact space of lower dimensionality. $mlp_{\mathcal{C}}()$ takes as input the new representations of v and C and outputs a score for the contextual relevance of v to C.

As training data, couples (v,C) from the BNC corpus are provided to the learning scheme. The target training value for the output of the system is 1 if v is indeed in C and -1 otherwise. The hope is that the neural network will be able to generalize to words which are not in the piece of text but are likely to be related to it.

In essence, this model is trained by minimizing the weighted sum of the hinge loss function over negative and positive couples, using stochastic Gradient Descent (see (Keller and Bengio, 2005) for further details). The small held out development set of the substitution dataset was used to tune the hyper-parameters of the model, maximizing average precision (see Section 4.1). For simplicity $mlp_{\mathcal{A}}()$ and $mlp_{\mathcal{B}}()$ were reduced to Perceptrons. The output size of $mlp_{\mathcal{A}}()$ was set to 20, $mlp_{\mathcal{B}}()$ to 100 and the number of hidden units of $mlp_{\mathcal{C}}()$ was set to 500.

There are a couple of important conceptual differences of the discriminative NNTR model compared to the generative Bayesian model described above. First, the relevancy of v to C in NNTR is inferred in a more compact representation space of reduced dimensionality, which may enable a higher degree of generalization. Second, in NNTR we are able to control the capacity of the model in terms of number of parameters, enabling better control to achieve an optimal generalization level with respect to the training data (avoiding over or under fitting).

4 Empirical Results

4.1 Evaluation Measures

We compare the lexical substitution scoring methods using two evaluation measures, offering two different perspectives of evaluation.

4.1.1 Accuracy

The first evaluation measure is motivated by simulating a decision step of a subtitling system, in which the best scoring lexical substitution is selected for each given sentence. Such decision may correspond to a situation in which each single substitution may suffice to obtain the desired compression rate, or might be part of a more complex decision mechanism of the complete subtitling system. We thus measure the resulting accuracy of subtitles created by applying the best scoring substitution example for every original sentence. This provides a macro evaluation style since we obtain a single judgment for each group of substitution examples that correspond to one original sentence.

In our dataset 25.5% of the original sentences have no correct substitution examples and for 15.5% of the sentences all substitution examples were annotated as correct. Accordingly, the (macro averaged) accuracy has a lower bound of 0.155 and upper bound of 0.745.

4.1.2 Average Precision

As a second evaluation measure we compare the *average precision* of each method over all the examples from all original sentences pooled together (a micro averaging approach). This measures the potential of a scoring method to ensure high precision for the high scoring examples and to filter out low-scoring incorrect substitutions.

Average precision is a single figure measure commonly used to evaluate a system's ranking ability (Voorhees and Harman, 1999). It is equivalent to the area under the uninterpolated recall-precision curve, defined as follows:

$$\text{average precision} = \frac{\sum_{i=1}^{N} \mathrm{P}(i) T(i)}{\sum_{i=1}^{N} T(i)} \qquad \mathrm{P}(i) = \frac{\sum_{k=1}^{i} T(k)}{i} \tag{2}$$

where N is the number of examples in the test set (797 in our case), $T(i)$ is the gold annotation (true=1, false=0) and i ranges over the examples ranked by decreasing score. An average precision of 1.0 means that the system assigned a higher score to all true examples than to any false one (perfect ranking). A lower bound of 0.26 on our test set corresponds to a system that ranks all false examples above the true ones.

4.2 Results

Figure 1 shows the accuracy and average precision results of the various models on our test set. The random baseline and corresponding significance levels were achieved by averaging multiple runs of a system that assigned random scores. As can be seen in the figures, the models' behavior seems to be consistent in both evaluation measures.

Overall, the distributional similarity based method (sim) performs much better than the other methods. In particular, Lin's similarity also performs better than semcor, the other context-independent model. Generally, the context independent models perform better than the contextual ones. Between the two contextual models, nntr is superior to Bayes. In fact the Bayes model is not significantly better than random scoring.

4.3 Analysis and Discussion

When analyzing the data we identified several reasons why some of the WordNet substitutions were judged as false. In some cases the source word as appearing in the original sentence is not in a sense for which it is a synonym of the target word. For example, in many situations the word *answer* is in the sense of a statement that is made in reply to a question or request. In such cases, such as in example 2 from Table 1, *answer* can be successfully replaced with *reply* yielding a substitution which conveys the original meaning. However, in situations such as in example 1 the word *answer* is in the sense of a general solution and cannot be replaced with *reply*. This is also the case in examples 4 and 5 in which *subject* does not appear in the sense of topic or theme.

Having an inappropriate sense, however, is not the only reason for incorrect substitutions. In example 8 *approach* appears in a sense which is synonymous with attack and in example 9 *problem* appears in a sense which is synonymous with a quite uncommon use of the word job. Nevertheless, these substitutions were judged as unacceptable since the desired sense of the target word after the substitution is not very clear from the context. In many other cases, such as in example 7, though semantically correct, the substitution was judged as incorrect due to stylistic considerations.

Finally, there are cases, such as in example 6 in which the source word is part of a collocation and cannot be replaced with semantically equivalent words.

When analyzing the mistakes of the distributional similarity method it seems as if many were not necessarily due to the method itself but rather to implementation issues. The online source we used contains only the top most similar words for any word. In many cases substitutions were assigned a score of zero since they were not listed among the top scoring similar words in the database. Furthermore, the corpus that was used for training the similarity scores was news articles in American English spelling and does not always supply good scores to words of British spelling in our BBC dataset (e.g. analyse, behavioural, etc.).

The similarity based method seems to perform better than the SemCor based method since, as noted above, even when the source word is in the appropriate sense it not necessarily substitutable with the target. For this reason we hypothesize that applying Word Sense Disambiguation (WSD) methods to

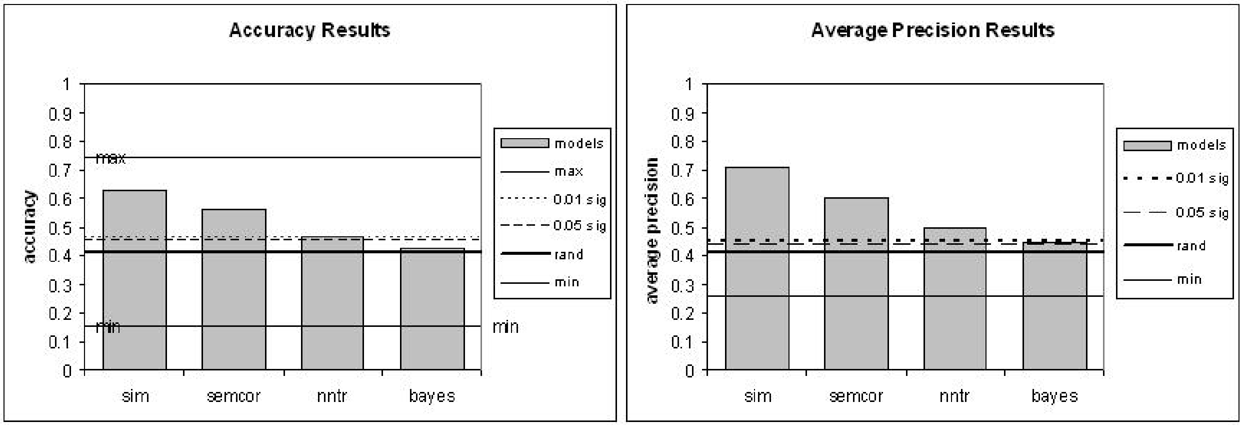

Figure 1: Accuracy and Average Precision Results

classify the specific WordNet sense of the source and target words may have only a limited impact on performance.

Overall, context independent models seem to perform relatively well since many candidate synonyms are a priori not substitutable. This demonstrates that such models are able to filter out many quirky WordNet synonyms, such as problem and job.

Fitness to the sentence context seems to be a less frequent factor and not that trivial to model. Local context (adjacent words) seems to play more of a role than the broader sentence context. However, these two types of contexts were not distinguished in the bag-of-words representations of the two contextual methods that we examined. It will be interesting to investigate in future research using different feature types for local and global context, as commonly done for Word Sense Disambiguation (WSD). Yet, it would still remain a challenging task to correctly distinguish, for example, the contexts for which *answer* is substitutable by *reply* (as in example 2) from contexts in which it is not (as in example 1).

So far we have investigated separately the performance of context independent and contextual models. In fact, the accuracy performance of the (context independent) sim method is not that far from the upper bound, and the analysis above indicated a rather small potential for improvement by incorporating information from a contextual method. Yet, there is still a substantial room for improvement in the ranking quality of this model, as measured by average precision, and it is possible that a smart combination with a high-quality contextual model would yield better performance. In particular, we would expect that a good contextual model will identify the cases in which for potentially good synonyms pair, the source word appears in a sense that is not substitutable with the target, such as in examples 1, 4 and 5 in Table 1. Investigating better contextual models and their optimal combination with context independent models remains a topic for future research.

5 Conclusion

This paper investigated an isolated setting of the lexical substitution task, which has typically been embedded in larger systems and not evaluated directly. The setting allowed us to analyze different types of state of the art models and their behavior with respect to characteristic sub-cases of the problem.

The major conclusion that seems to arise from our experiments is the effectiveness of combining a knowledge based thesaurus such as WordNet with distributional statistical information such as (Lin, 1998), overcoming the known deficiencies of each method alone. Furthermore, modeling the a priori substitution likelihood captures the majority of cases in the evaluated setting, mostly because WordNet provides a rather noisy set of substitution candidates. On the other hand, successfully incorporating local and global contextual information, as similar to WSD methods, remains a challenging task for future research. Overall, scoring lexical substitutions is an important component in many applications and we expect that our findings are likely to be broadly applicable.

References

[Budanitsky and Hirst2001] Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources: Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 29–34.

[Burnard1995] Lou Burnard. 1995. *Users Reference Guide for the British National Corpus*. Oxford University Computing Services, Oxford.

[Daelemans et al.2004] Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

[Dagan et al.2005] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment.

[Geffet and Dagan2005] Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114, Ann Arbor, Michigan, June. Association for Computational Linguistics.

[Glickman et al.2005] Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. A probabilistic classification approach for lexical textual entailment. In *AAAI*, pages 1050–1055.

[Grefenstette1998] Gregory Grefenstette. 1998. Producing Intelligent Telegraphic Text Reduction to Provide an Audio Scanning Service for the Blind. pages 111–117, Stanford, CA, March.

[Harris1968] Zelig Harris. 1968. *Mathematical Structures of Language*. New York: Wiley.

[Hori et al.2002] Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex Waibel. 2002. Automatic speech summarization applied to english broadcast news speech. volume 1, pages 9–12.

[Jing and McKeown1999] Hongyan Jing and Kathleen R. McKeown. 1999. The decomposition of human-written summary sentences. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136, New York, NY, USA. ACM Press.

[Keller and Bengio2005] Mikaela Keller and Samy Bengio. 2005. A neural network for text representation. In Wodzisaw Duch, Janusz Kacprzyk, and Erkki Oja, editors, *Artificial Neural Networks: Biological Inspirations ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part II*, volume 3697 / 2005 of *Lecture Notes in Computer Science*, page p. 667. Springer-Verlag GmbH.

[Knight and Marcu2002] Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artif. Intell.*, 139(1):91–107.

[Landis and Koch1997] J. R. Landis and G. G. Koch. 1997. The measurements of observer agreement for categorical data. *Biometrics*, 33:159–174.

[Lin1998] Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.

[McCarthy et al.2004] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *ACL*, pages 280–288, Morristown, NJ, USA. Association for Computational Linguistics.

[Miller et al.1993] George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 303–308, Morristown, NJ, USA. Association for Computational Linguistics.

[Piperidis et al.2004] Stelios Piperidis, Iason Demiros, Prokopis Prokopidis, Peter Vanroose, Anja Höthker, Walter Daelemans, Elsa Sklavounou, Manos Konstantinou, and Yannis Karavidas. 2004. Multimodal multilingual resources in the subtitling process. In *Proceedings of the 4th International Language Resources and Evaluation Conference (LREC 2004)*, Lisbon.

[Voorhees and Harman1999] Ellen M. Voorhees and Donna Harman. 1999. Overview of the seventh text retrieval conference. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication.

[Voorhees1994] Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.