Lexical Reference: a Semantic Matching Subtask

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

Abstract

Semantic lexical matching is a prominent subtask within text understanding applications. Yet, it is rarely evaluated in a direct manner. This paper proposes a definition for *lexical reference* which captures the common goals of lexical matching. Based on this definition we created and analyzed a test dataset that was utilized to directly evaluate, compare and improve lexical matching models. We suggest that such decomposition of the global semantic matching task is critical in order to fully understand and improve individual components.

1 Introduction

A fundamental task for text understanding applications is to identify semantically equivalent pieces of text. For example, Question Answering (QA) systems need to match corresponding parts in the question and in the answer passage, even though such parts may be expressed in different terms. Summarization systems need to recognize (redundant) semantically matching parts in multiple sentences that are phrased differently. Other applications, such as information extraction and retrieval, face pretty much the same semantic matching task. The degree of semantic matching found is typically factored into systems' scoring and ranking mechanisms. The recently proposed framework of textual entailment (Dagan et al., 2006) attempts to formulate the generic semantic matching problem in an application independent manner.

The most commonly implemented semantic matching component addresses the lexical level. At this level the goal is to identify whether the meaning of a lexical item of one text is expressed also within the other text. Typically, lexical matching models measure the degree of literal lexical overlap, augmented with lexical substitution criteria based on resources such as Wordnet or the output of statistical similarity methods (see Section 2). Many systems apply semantic matching only at the lexical level, which is used to approximate the overall degree of semantic matching between texts. Other systems incorporate lexical matching as a component within more complex models that examine matching at higher syntactic and semantic levels.

While lexical matching models are so prominent within semantic systems they are rarely evaluated in a direct manner. Typically, improvements to a lexical matching model are evaluated by their marginal contribution to overall system performance. Yet, such global and indirect evaluation does not indicate the absolute performance of the model relative to the sheer lexical matching task for which it was designed. Furthermore, the indirect application-dependent evaluation mode does not facilitate improving lexical matching models in an application dependent manner, and does not allow proper comparison of such models which were developed (and evaluated) by different researchers within different systems.

This paper proposes a generic definition for the lexical matching task, which we term *lexical reference*. This definition is application independent and enables annotating test datasets that evaluate directly lexical matching models. Consequently, we created a dataset annotated for lexical reference, using a sample of sentence pairs (text-hypothesis) from the 1st Recognising Textual Entailment dataset. Further analysis identified sev-

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 172–179,
Sydney, July 2006.

eral sub-types of lexical reference, pointing at the many interesting cases where lexical reference is derived from a complete context rather than from a particular matching lexical item.

Next, we used the lexical reference dataset to evaluate and compare several state-of-the-art approaches for lexical matching. Having a direct evaluation task enabled us to capture the actual performance level of these models, to reveal their relative strengths and weaknesses, and even to construct a simple combination of two models that outperforms all the original ones. Overall, we suggest that it is essential to decompose global semantic matching and textual entailment tasks into proper subtasks, like lexical reference. Such decomposition is needed in order to fully understand the behavior of individual system components and to guide their future improvements.

2 Background

2.1 Term Matching

Thesaurus-based term expansion is a commonly used technique for enhancing the recall of NLP systems and coping with lexical variability. Expansion consists of altering a given text (usually a query) by adding terms of similar meaning. WordNet is commonly used as a source of related words for expansion. For example, many QA systems perform expansion in the retrieval phase using query related words based on WordNet's lexical relations such as synonymy or hyponymy (e.g (Harabagiu et al., 2000; Hovy et al., 2001)). Lexical similarity measures (e.g. (Lin, 1998)) have also been suggested to measure semantic similarity. They are based on the distributional hypothesis, suggesting that words that occur within similar contexts are semantically similar.

2.2 Textual Entailment

The Recognising Textual Entailment (RTE-1) challenge (Dagan et al., 2006) is an attempt to promote an abstract generic task that captures major semantic inference needs across applications. The task requires to recognize, given two text fragments, whether the meaning of one text can be inferred (entailed) from another text. Different techniques and heuristics were applied on the RTE-1 dataset to specifically model textual entailment. Interestingly, a number of works (e.g. (Bos and Markert, 2005; Corley and Mihalcea, 2005; Jijkoun and de Rijke, 2005; Glickman et al., 2006)) applied or utilized lexical based word overlap measures. Various word-to-word similarity measures where applied, including distributional similarity (such as (Lin, 1998)), web-based co-occurrence statistics and WordNet based similarity measures (such as (Leacock et al., 1998)).

2.3 Paraphrase Acquisition

A substantial body of work has been dedicated to learning patterns of semantic equivalency between different language expressions, typically considered as paraphrases. Recently, several works addressed the task of acquiring paraphrases (semi-) automatically from corpora. Most attempts were based on identifying corresponding sentences in parallel or 'comparable' corpora, where each corpus is known to include texts that largely correspond to texts in another corpus (e.g. (Barzilay and McKeown, 2001)). Distributional Similarity was also used to identify paraphrase patterns from a single corpus rather than from a comparable set of corpora (Lin and Pantel, 2001). Similarly, (Glickman and Dagan, 2004) developed statistical methods that match verb paraphrases within a regular corpus.

3 The Lexical Reference Dataset

3.1 Motivation and Definition

One of the major observations of the 1st Recognizing Textual Entailment (RTE-1) challenge referred to the rich structure of entailment modeling systems and the need to evaluate and optimize individual components within them. When building such a compound system it is valuable to test each component directly during its development, rather than indirectly evaluating the component's performance via the behavior of the entire system. If given tools to evaluate each component independently researchers can target and perfect the performance of the subcomponents without the need of building and evaluating the entire end-to-end system.

A common subtask, addressed by practically all participating systems in RTE-1, was to recognize whether each lexical meaning in the hypothesis is referenced by some meaning in the corresponding text. We suggest that this common goal can be captured through the following definition:

Definition 1 *A word w is **lexically referenced** by a text t if there is an explicit or implied reference*

from a set of words in t to a possible meaning of w.

Lexical reference may be viewed as a natural extension of textual entailment for sub-sentential hypotheses such as words. In this work we focus on words meanings, however this work can be directly generalized to word compounds and phrases. A concrete version of detailed annotation guidelines for lexical reference is presented in the next section.[1] Lexical Reference is, in some sense, a more general notion than paraphrases. If the text includes a paraphrase for w then naturally it does refer to w's meaning. However, a text need not include a paraphrase for the concrete meaning of the referenced word w, but only an implied reference. Accordingly, the referring part might be a large segment of the text, which captures information different than w's meaning, but still implies a reference to w as part of the text's meaning.

It is typically a necessary, but not sufficient, condition for textual entailment that the lexical concepts in a hypothesis h are referred in a given text t. For example, in order to infer from a text the hypothesis "*a dog bit a man*," it is a necessary that the concepts of *dog*, *bite* and *man* must be referenced by the text, either directly or in an implied manner. However, for proper entailment it is further needed that the right relations would hold between these concepts[2]. Therefore lexical entailment should typically be a component within a more complex entailment modeling (or semantic matching) system.

3.2 Dataset Creation and Annotation Process

We created a lexical reference dataset derived from the RTE-1 development set by randomly choosing 400 out of the 567 text-hypothesis examples. We then created sentence-word examples for all content words in the hypotheses which do not appear in the corresponding sentence and are not a morphological derivation of a word in it (since a simple morphologic module could easily identify these cases). This resulted in a total of 708 lexical reference examples. Two annotators annotated these examples as described in the next section. Taking the same approach as of the RTE-1 dataset creation (Dagan et al., 2006), we limited our experiments to the resulting 580 examples that the two annotators agreed upon[3].

3.2.1 Annotation guidelines

We asked two annotators to annotate the sentence-word examples according to the following guidelines. Given a sentence and a target word the annotators were asked to decide whether the target word is referred by the sentence (true) or not (false). Annotators were guided to mark the pair as true in the following cases:

Word: if there is a word in the sentence which, in the context of the sentence, implies a meaning of the target word (e.g. a synonym or hyponym), or which implies a reference to the target word's meaning (e.g. blind→see, sight). See examples 1-2 in Table 1 where the word that implies the reference is emphasized in the text. Note that in example 2 murder is not a synonym of died nor does it share the same meaning of died; however it is clear from its presence in the sentence that it refers to a death. Also note that in example 8 although home is a possible synonym for house, in the context of the text it does not appear in that meaning and the example should be annotated as false.

Phrase: if there is a multi-word independent expression in the sentence that implies the target (implication in the same sense that a Word does). See examples 3-4 in Table 1.

Context: if there is a clear reference to the meaning of the target word by the overall meaning of some part(s) of the sentence (possibly all the sentence), though it is not referenced by any single word or phrase. The reference is derived from the complete context of the relevant sentence part. See examples 5-7 in Table 1.

If there is no reference from the sentence to the target word the annotators were instructed to choose false. In example 9 in Table 1 the target word "HIV-positive" should be considered as one word that cannot be broken down from its unit and although both the general term "HIV status" and the more specific term "HIV negative" are referred to, the target word cannot be understood or derived from the text. In example 10 although the year 1945 may refer to a specific war, there is no "war" either specifically or generally understood by the text.

[1] These terms should not be confused with the use of *lexical entailment* in WordNet, which is used to describe an entailment relationship between verb lexical types, nor with the related notion of *reference* in classical linguistics, generally describing the relation between nouns or pronouns and objects that are named by them (Frege, 1892)

[2] or quoting the known journalism saying – "*Dog bites man*" isn't news, but "*Man bites dog*" is.

[3] dataset avaiable at `http://ir-srv.cs.biu.ac.il:64080/emnlp06_dataset.zip`

ID	TEXT	TARGET	VALUE
1	Oracle had fought to keep the **forms** from being released.	document	word
2	The court found two men guilty of **murdering** Shapour Bakhtiar.	died	word
3	The new information prompted them to **call off** the search.	cancelled	phrase
4	Milan, **home of** the famed La Scala opera house,...	located	phrase
5	Successful **plaintiffs** recovered **punitive damages** in Texas discrimination **cases** 53	legal	context
6	Recreational marijuana smokers are no more **likely to develop oral cancer** than nonusers.	risk	context
7	A bus ticket **cost nowadays 5.2 NIS whereas last year it cost 4.9**.	increase	context
8	Pakistani officials announced that two South African men in their custody had confessed to planning attacks at popular tourist spots in their home country.	house	false
9	For women who are HIV negative or who do not know their HIV status, breastfeeding should be promoted for six months.	HIV-positive	false
10	On Feb. 1, 1945, the Polish government made Warsaw its capital, and an office for urban reconstruction was set up.	war	false

Table 1: Lexical Reference Annotation Examples

3.2.2 Annotation results

We measured the agreement on the lexical reference binary task (in which Word, Phrase and Context are conflated to true). The resulting kappa statistic of 0.63 is regarded as substantial agreement (Landis and Koch, 1997). The resulting dataset is not balanced in terms of true and false examples and a straw-baseline for accuracy is 0.61, representing a system which predicts all examples as true.

3.3 Dataset Analysis

In a similar manner to (Bar-Haim et al., 2005; Vanderwende et al., 2005) we investigated the relationship between lexical reference and textual entailment. We checked the performance of a textual entailment system which relies solely on an ideal lexical reference component which makes no mistakes and asserts that a hypothesis is entailed from a text if and only if all content words in the hypothesis are referred in the text. Based on the lexical reference dataset annotations, such an "ideal" system would obtain an accuracy of 74% on the corresponding subset of the textual entailment task. The corresponding precision is 68% and a recall of 82%. This is significantly higher than the results of the best performing systems that participated in the challenge on the RTE-1 test set. This suggests that lexical reference is a valuable subtask for entailment. Interestingly, a similar entailment system based on a lexical reference component which doesn't account for the contextual lexical reference (i.e. all Context annotations are regarded as false) would achieve an accuracy of only 63% with 41% precision and a recall of 63%. This suggests that lexical reference in general and contextual entailment in particular, play an important (though not sufficient) role in entailment recognition.

Further, we wanted to investigate the validity of the assumption that for entailment relationship to hold all content words in the hypothesis must be referred by the text. We examined the examples in our dataset which were derived from text-hypothesis pairs that were annotated as true (entailing) in the RTE dataset. Out of 257 such examples only 34 were annotated as false by both annotators. Table 2 lists a few such examples in which entailment at whole holds, however, there exists a word in the hypothesis (highlighted in the table) which is not lexically referenced by the text. In many cases, the target word was part of a non compositional compound in the hypothesis, and therefore should not be expected to be referenced by the text (see examples 1-2). This finding indicates that the basic assumption is a reasonable approximation for entailment. We could not have revealed this fact without the dataset for the subtask of lexical reference.

4 Lexical Reference Models

The lexical reference dataset facilitates qualitative and quantitative comparison of various lexical models. This section describes four state-of-the-art models that can be applied to the lexical reference task. The performance of these models was tested and analyzed, as described in the next section, using the lexical reference dataset. All models assign a $[0, 1]$ score to a given pair of text t and target word u which can be interpreted as the confidence that u is lexically referenced in t.

ID	TEXT	HYPOTHESIS	ENTAIL-MENT	REFER-ENCE
1	Iran is said to give up al Qaeda members.	Iran **hands** over al Qaeda members.	true	false
2	It would help the economy by putting people back to work and more money in the hands of consumers.	More money in the hands of consumers means more money can be spent to get the economy **going**.	true	false
3	The Securities and Exchange Commission's new rule to beef up the independence of mutual fund boards represents an industry defeat.	The SEC's new rule will **give** boards independence.	true	false
4	Texas Data Recovery is also successful at retrieving lost data from notebooks and laptops, regardless of age, make or model.	In the event of a **disaster** you could use Texas Data Recovery and you will have the capability to restore lost data.	true	false

Table 2: examples demonstrating cases when lexical entailment does not correlate with entailment. Target word is shown in bold.

4.1 WordNet

Following the common practice in NLP applications (see Section 2.1) we evaluated the performance of a straight-forward utilization of WordNet's lexical information. Our wordnet model first lemmatizes the text and target word. It then assigns a score of 1 if the text contains a synonym, hyponym or derived form of the target word and a score of 0 otherwise.

4.2 Similarity

As a second measure we used the distributional similarity measure of (Lin, 1998). For a text t and a word u we assign the max similarity score as follows:

$$similarity(t,u) = \max_{v \in t} sim(u,v) \qquad (1)$$

where $sim(u,v)$ is the similarity score for u and v[4].

4.3 Alignment model

(Glickman et al., 2006) was among the top scoring systems on the RTE-1 challenge and supplies a probabilistically motivated lexical measure based on word co-occurrence statistics. It is defined for a text t and a word u as follows:

$$align(t,u) = \max_{v \in t} \mathrm{P}(u|v) \qquad (2)$$

where $\mathrm{P}(u|v)$ is simply the co-occurrence probability – the probability that a sentence containing v also contains u. The co-occurrence statistics were collected from the Reuters Corpus Volume 1.

[4]the scores were obtained from the following online resource: `http://www.cs.ualberta.ca/~lindek/downloads.htm`

4.4 Baysean model

(Glickman et al., 2005) provide a contextual measure which takes into account the whole context of the text rather than from a single word in the text as do the previous models. This model is the only model which addresses contextual reference rather than just word-to-word matching. The model is based on a Naïve Bayes text classification approach in which corpus sentences serve as documents and the class is the reference of the target word u. Sentences containing the word u are used as positive examples while all other sentences are considered as negative examples. It is defined for a text t and a word u as follows:

$$bayes(t,u) = \frac{\mathrm{P}(u)\prod_{v \in t}\mathrm{P}(v|u)^{n(v,t)}}{\mathrm{P}(\neg u)\prod_{v \in t}\mathrm{P}(v|\neg u)^{n(v,t)} + \mathrm{P}(u)\prod_{v \in t}\mathrm{P}(v|u)^{n(v,t)}} \qquad (3)$$

where $n(w,t)$ is the number of times word w appears in t, $\mathrm{P}(u)$ is the probability that a sentence contains the word u and $\mathrm{P}(v|\neg u)$ is the probability that a sentence NOT containing u contains v. In order to reduce data size and to account for zero probabilities we applied smoothing and information gain based feature selection on the data prior to running the model. The co-occurrence probabilities were collected from sentences from the Reuters corpus in a similar manner to the alignment model.

4.5 Combined Model

The WordNet and Bayesian models are derived from quite different motivations. One would expect the WordNet model to be better in identifying the word-to-word explicit reference examples while the Bayesian model is expected to model the contextualy implied references. For this reason we tried to combine forces by evaluating a naïve linear

interpolation of the two models (by simply averaging the score of the two models). This model have not been previously suggested and to the best of our knowledge this type of combination is novel.

5 Empirical Evaluation and Analysis

5.1 Results

In order to evaluate the scores produced by the various models as a potential component in an entailment system we compared the recall-precision graphs. In addition we compared the *average precision* which is a single number measure equivalent to the area under an uninterpolated recall-precision curve and is commonly used to evaluate a systems ranking ability (Voorhees and Harman, 1999). On our dataset an average precision greater than 0.65 is better than chance at the 0.05 level and an average precision greater than 0.66 is significant at the 0.01 level.

Figure 1 compares the average precision and recall-precision results for the various models. As can be seen, the combined wordnet+bayes model performs best. In terms of average precision, the similarity and wordnet models are comparable and are slightly better than bayes. The alignment model, however, is not significantly better than random guessing. The recall-precision figure indicates that the baysian model succeeds to rank quite well both within the the positively scored wordnet examples and within the negatively scored wordnet examples and thus resulting in improved average precision of the combined model. A better understanding of the systems' performance is evident from the following analysis.

5.2 Analysis

Table 3 lists a few examples from the lexical reference dataset along with their gold-standard annotation and the Bayesian model score. Manual inspection of the data shows that the Bayesian model commonly assigns a low score to correct examples which have an entailing trigger word or phrase in the sentence but yet the context of the sentence as a whole is not typical for the target hypothesized entailed word. For example, in example 5 the entailing phrase 'set in place' and in example 6 the entailing word 'founder' do appear in the text however the contexts of the sentences are not typical news domain contexts of issued or founded. An interesting future work would be to change the generative story and model to account for such cases.

The WordNet model identified a matching word in the text for 99 out of the 580 examples. This corresponds to a somewhat low recall of 25% and a quite high precision of 90%. Table 4 lists typical mistakes of the wordnet model. Examples 1-3 are false positive examples in which there is a word in the text (emphasized in the table) which is a synonym or hyponym of the target word for some sense in WordNet, however in the context of the text it is not of such a sense. Examples 4-6 show false negative examples, in which the annotators identified a trigger word in the text (emphasized in the table) but yet it or no other word in the text is a synonym or hyponym of the target word.

5.3 Subcategory analysis

	word	phrase	context	false
word	178	16	59	32
phrase	4	12	9	4
context	15	5	56	25
false	24	5	38	226

Table 5: inter-annotator confusion matrix for the auxiliary annotation.

As seen above, the combined model outperforms the others since it identifies both word-to-word lexical reference as well as context-to-word lexical reference. These are quite different cases. We asked the annotators to state the subcategory when they annotated an example as true (as described in the annotation guidelines in Section 3.2.1). The *Word* subcategory corresponds to a word-to-word match and *Phrase* and *Context* subcategories correspond to more than one word to word match. As can be expected, the agreement on such a task resulted in a lower Kappa of 0.5 which corresponds to moderate agreement (Landis and Koch, 1997). the confusion matrix between the two annotators is presented in Table 5. This decomposition enables the evaluation of the strength and weakness of different lexical reference modules, free from the context of the bigger entailment system.

We used the subcategories dataset to test the performances of the different models. Table 6 lists for each subcategory the recall of correctly identified examples for each model's 25% recall level. The table shows that the wordnet and similarity models' strength is in identifying examples where lexical reference is triggered by a dominant word in the sentence. The bayes model, however,

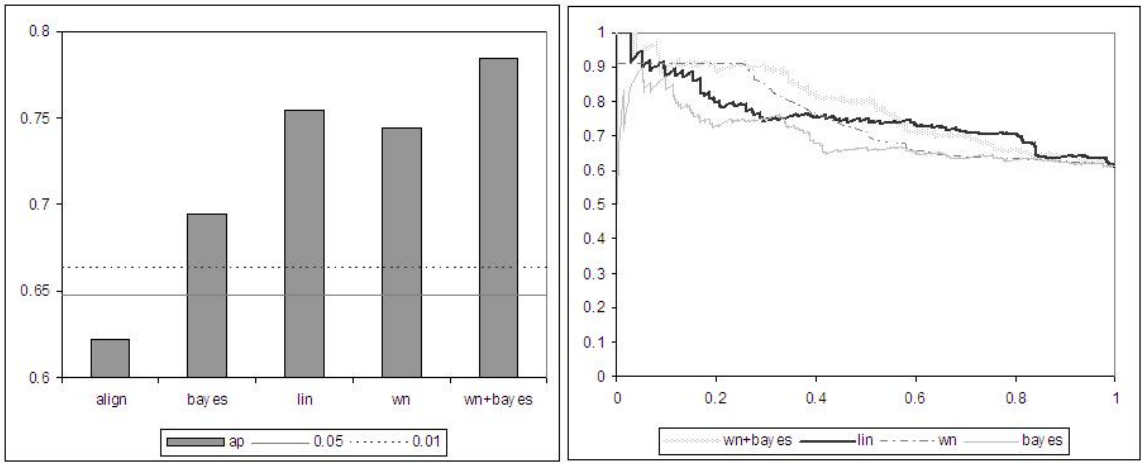

Figure 1: comparison of average precision (left) and recall-precision (right) results for the various models

id	text	token	annotation	score
1	*QNX Software Systems Ltd., a leading provider of real-time software and services to the embedded computing market, is pleased to announce the appointment of Mr. Sachin Lawande to the position of vice president, engineering services.*	named	PHRASE	0.98
2	*NIH's FY05 budget request of $28.8 billion includes $2 billion for the National Institute of General Medical Sciences, a 3.4-percent increase, and $1.1 billion for the National Center for Research Resources, and a 7.2-percent decrease from FY04 levels.*	reduced	WORD	0.91
3	*Pakistani officials announced that two South African men in their custody had confessed to planning attacks at popular tourist spots in their home country.*	security	CONTEXT	0.80
4	*With $549 million in cash as of June 30, Google can easily afford to make amends.*	shares	FALSE	0.03
5	*In the year 538, Cyrus set in place a policy which demanded the return of the various gods to their proper places.*	issued	PHRASE	7e-4
6	*The black Muslim activist said that he had relieved Muhammad of his duties "until he demonstrates that he is willing to conform to the manner of representing Allah and the honorable Elijah Muhammad (founder of the Nation of Islam)".*	founded	WORD	3e-6

Table 3: A sample from the lexical reference dataset along with the Bayesian model's score

id	text	token	annotation
1	*Kerry **hit** Bush hard on his conduct on the war in Iraq*	shot	FALSE
2	*Pakistani officials announced that two South African **men** in their custody had confessed to planning attacks at popular tourist spots in their home country*	forces	FALSE
3	*It would help the economy by putting people back to **work** and more money in the hands of consumers*	get	FALSE
4	*Eating lots of foods that are a good source of fiber may keep your blood **glucose** from rising too fast after you eat*	sugar	WORD
5	*Hippos do come into conflict with **people** quite often*	human	WORD
6	*Weinstock painstakingly reviewed dozens of studies for evidence of any link between sunscreen use and either an increase or decrease in **melanoma***	cancer	WORD

Table 4: A few erroneous examples of WordNet model

is better at identifying phrase and context examples. The combined WordNet and Bayesian models' strength can be explained by the quite different behaviors of the two models - the WordNet model seems to be better in identifying the word-to-word explicit reference examples while the Bayesian model is better in modeling the contextual implied references.

6 Conclusions

This paper proposed an explicit task definition for *lexical reference*. This task captures directly the goal of common lexical matching models, which typically operate within more complex systems

method	word	disagreement	phrase/context
wordnet	38%	9%	17%
similarity	39%	7%	17%
bayes	22%	21%	37%

Table 6: Breakdown of recall of correctly identified example types at an overall system's recall of 25%. Disagreement refers to examples for which the annotators did not agree on the subcategory annotation (word vs. phrase/context).

that address more complex tasks. This definition enabled us to create an annotated dataset for the lexical reference task, which provided insights into interesting sub-classes that require different types of modeling. The dataset enabled us to make a direct evaluation and comparison of lexical matching models, reveal insightful differences between them, and create a simple improved model combination. In the long run, we believe that the availability of such datasets will facilitate improved models that consider the various sub-cases of lexical reference, as well as applying supervised learning to optimize model combination and performance.

References

[Bar-Haim et al.2005] Roy Bar-Haim, Idan Szpecktor, and Oren Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 55–60, Ann Arbor, Michigan, June. Association for Computational Linguistics.

[Barzilay and McKeown2001] Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *ACL*, pages 50–57.

[Bos and Markert2005] Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference techniques. In *EMNLP*.

[Corley and Mihalcea2005] Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18.

[Dagan et al.2006] Ido Dagan, Oren Glickman, and Bernardo Magnini, editors. 2006. *The PASCAL Recognising Textual Entailment Challenge*, volume 3944. Lecture Notes in Computer Science.

[Frege1892] Gottlob Frege. 1892. On sense and reference. Reprinted in P. Geach and M. Black, eds., Translations from the Philosophical Writings of Gottlob Frege. 1960.

[Glickman and Dagan2004] Oren Glickman and Ido Dagan, 2004. *Recent Advances in Natural Language Processing III*, chapter Acquiring lexical paraphrases from a single corpus, pages 81–90. John Benjamins.

[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.

[Glickman et al.2006] Oren Glickman, Ido Dagan, and Moshe Koppel. 2006. A lexical alignment model for probabilistic textual entailment, volume 3944. In *Lecture Notes in Computer Science*, pages 287 – 298. Springer.

[Harabagiu et al.2000] Sanda M. Harabagiu, Dan I. Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C. Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. Falcon: Boosting knowledge for answer engines. In *TREC*.

[Hovy et al.2001] Eduard H. Hovy, Ulf Hermjakob, and Chin-Yew Lin. 2001. The use of external knowledge of factoid QA. In *Text REtrieval Conference*.

[Jijkoun and de Rijke2005] Valentin Jijkoun and Maarten de Rijke. 2005. Recognizing textual entailment using lexical similarity. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (and forthcoming LNAI book chapter).

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

[Leacock et al.1998] Claudia Leacock, George A. Miller, and Martin Chodorow. 1998. Using corpus statistics and wordnet relations for sense identification. *Comput. Linguist.*, 24(1):147–165.

[Lin and Pantel2001] Dekang Lin and Patrik Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 4(7):343–360.

[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.

[Vanderwende et al.2005] Lucy Vanderwende, Deborah Coughlin, and Bill Dolan. 2005. What syntax can contribute in entailment task. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment.

[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.