Artif Intell Rev (2006) 25:55–65
DOI 10.1007/s10462-007-9018-5

Identifying semantic equivalence for multi-document summarisation

Eamonn Newman · Joe Carthy · John Dunnion · Nicola Stokes

Published online: 22 August 2007

Abstract We describe Semantic Equivalence and Textual Entailment Recognition, and outline a system which uses a number of lexical, syntactic and semantic features to classify pairs of sentences as "semantically equivalent". We describe an experiment to show how syntactic and semantic features improve the performance of an earlier system, which used only lexical features. We also outline some areas for future work.

Keywords Semantic equivalence · Textual entailment · Document summarisation · NLP

1 Recognising semantic equivalence and textual entailment

Our current research focuses on the problem of Recognising Semantic Equivalence and Textual Entailment (RSETE). Two sentences are semantically equivalent if they attempt to convey the same information, or if one is entailed by the other, i.e., the information conveyed by one is covered by the information conveyed by the other.

An RSETE system could be used to improve the performance in many Natural Language Processing applications and domains. We outline some ways in which this could be achieved in this section.

This research originally stemmed from our work in the investigation of means of summarising text documents in general, with a specific focus on accident reports and other documents used in Incident and Accident Analysis. In (multi-document) summarisation, the goal is to convey in a summary the most salient information in the source document(s).

E. Newman (✉) · J. Carthy · J. Dunnion
School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin 4, Ireland
e-mail: eamonn.newman@ucd.ie

N. Stokes
NICTA Victoria Laboratory, Department of Computer Science and Software Engineering,
University of Melbourne, Melbourne, Australia

Currently, this is most commonly achieved using an extractive approach, in which the summariser calculates the most important parts of a document and uses them to build the summary.

The problem with this approach is that the most important information in a piece is liable to be repeated a number of times in the original (possibly in different synonymous forms). Thus a summary may contain a number of sentences which cover the same semantic information.

This motivates our interest in devising a means to detect this semantic overlap or semantic equivalence. A system using this would allow us to eliminate information repetition in summaries, thus improving the summary quality (information as a proportion of the length).

On further investigation, we came to see that a semantic equivalence system could be used to great benefit in other areas of research.

In question–answering, a system is given a question to which it must find the answer in a prescribed document or set of documents. We can use RSETE in this domain in quite a simple application by taking the question and reformulating it as a statement (possibly ungrammatically). We can then use the reformulation as a text and attempt to find a piece of text within the document which is either semantically equivalent or contradictory. The result of this search will be the answer to the original question.

More recently, QA testbeds have moved away from the simpler style questions which were generally looking for a statement of fact to serve as an answer. Research is now being directed towards more subjective questions which require a somewhat "deeper" approach. These questions require the system to make a high-level agglomeration of facts (often presented in a number of documents). While earlier systems could operate on simple word- or pattern-matching practices, now systems require an element of "comprehension" to perform. It remains to be seen if RSETE can meet the demands of this new generation of QA corpora, but it certainly seems worthy of investigation.

In Machine Translation, the quality of a translation (and thus the MT system which produced it) is based, in general terms, on a word-for-word comparison with some gold standard. Thus, an MT system which uses a different vocabulary may provide a very good translation yet score poorly due to a lack of word matches with the gold standard. A RSETE system would allow for a vocabulary-independent comparison of the translations, free of any bias of word choice or sentence composition.

In Information Extraction, state-of-the-art systems rely heavily on pattern-recognition and pattern-matching. With a system which can identify semantic equivalence, information can be extracted using one pattern on a number of semantically-equivalent texts.

Semantic Equivalence can be leveraged in Information Retrieval for Query Expansion. Generally Query Expansion is handled by taking query terms and the most significant terms for the retrieved documents, and combining these to make a new query. However this can suffer disadvantages of word-sense ambiguity. This could be avoided by using semantic equivalence to ensure all terms are semantically related to the original query terms.

In Sect. 2 we compare some of the current research in the area of Semantic Equivalence detection. In Sect. 3, we provide a thorough description of our system, and the features used to detect Textual Entailment. Section 4 outlines the corpora we used for training and testing our classifiers. We also describe the nature of the experiments conducted. The results of these experiments are presented in Sect. 4.4. Section 5 contains some examples of errors made by our system and a discussion of aspects of the system's performance which warrant further research. Finally, in Sect. 6, we draw some conclusions of our work thus far and outline areas of planned future work.

2 State of the art in semantic equivalence recognition

A number of different approaches have been taken in attempts to solve the problem of recognising Semantic Equivalence. In this section, we will describe some of the current systems and the theory behind them.

Many of the RSETE systems use some sort of lexical comparison, ranging from bag-of-words matching to cosine similarity to n-gram and longest common substring.

Many of the existing systems try to tackle the problem from the aspect of logical equivalence, translating the text to a number of theorems. If the system can prove that the theorems representing one sentence can be deduced from another, then the system can say that one sentence is logically entailed by another. While this is not strictly the same as semantic equivalence, there is a strong correlation between the two tasks.

Other systems use techniques taken from Machine Translation. Here, the driving force is that two instances are deemed to come from two different languages and if a translation can be found from one to the other, then we can consider the two texts to be semantically equivalent.

There are a variety of approaches that can be used to address the problem of recognising semantic equivalence, as is evident by the breadth of the applications presented at the PASCAL RTE workshop (Dagan et al. 2005) and the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment (Dolan and Dagan 2005). Most of these systems use some sort of lexical matching, be it simple word overlap or some more complex statistical co-occurrence relation (e.g. Latent Semantic Indexing). While these systems perform better than those without lexical matching, it was widely agreed that matching at a word-level alone was not sufficient for the PASCAL corpus. Corley and Mihalcea (2005) presents an overview of similarity metrics based on WordNet concepts. They showed that a combination of WordNet similarity measures (Budanitsky and Hirst 2001) with a lexical matching metric (based on the number of shared words in a sentence-pair) achieved scores on the PASCAL corpus of up to 58.9%, which is comparable with other high-ranking systems at the workshop.

A number of the systems (de Salvo Braz et al. 2005; Akhmatova 2005; Bos and Markert 2005) used logical inference in which a representation of the text and hypothesis is constructed, and then a proof of the hypothesis is derived for the text (some of these systems appealed to world knowledge (hand-coded (Fowler et al. 2005); geographical Bos and Markert 2005), or to formal lexical resources such as WordNet).

A number of systems represented the texts as parse trees (e.g. syntactic, dependency, semantic) (Pazienza et al. 2005; Herrera et al. 2005). This action reduces the problem of textual entailment recognition to one of (sub–)graph matching.

Interestingly, Vanderwende et al. (2005) showed that using no more than syntactic matching, one could match up to 37% of classifications correctly. Appealing to a thesaurus yields up to 49%. This is supported by empirical evidence from Herrera et al. (2005) and Marsi and Krahmer (2005). Hence, it seems that relatively simple metrics used in combination perform better than more complex, "deeper" metrics such as logical inference or the incorporation of world knowledge into the classification computation. We suggest that this is the case because deep linguistic and inferential analysis is more prone to errors due to problems arising from word sense disambiguation.

One of the top systems in the PASCAL evaluation (Raina et al. 2005a, b) used all of the methods outlined above to some degree. Parsed sentences are represented as logical formulae. A theorem prover is then used to find the minimum cost of "proving" that the hypothesis is entailed by the text. These costs are learned from syntactic and semantic features and resources such as WordNet.

Table 1 Features used by decision-tree classifier

`entails`	Boolean, unknown
`<rouge>`	Continuous
`cosine`	Continuous
`LSI`	Continuous
`<wordnet>`	Continuous
`<VerbOcean>`	Continuous
`negation_t`	Continuous
`negation_h`	Continuous
`negdiff`	Continuous
`<lcs>`	Boolean
`lcs+not`	Boolean

<*name*> indicates a tuple of related features

3 System description

Our system uses a decision tree classifier whose features include lexical, semantic and grammatical attributes of nouns, verbs and adjectives to identify an entailment relationship between pairs of sentences (comprising of a *text* and a *hypothesis*). We generated our classifier from training sets using the C5.0 machine learning algorithm (Quinlan 2002).

The features used are calculated using the WordNet taxonomy (Fellbaum 1998), the VerbOcean semantic network (Chklovski and Pantel 2004) (developed at ISI) and a Latent Semantic Indexing (Deerwester et al. 1990) technique. Other features are based on the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy 2004) n-gram overlap metrics and cosine similarity between the text and hypothesis.

Our most sophisticated linguistic feature finds the longest common subsequence in the sentence-pair, and then detects contradictions in the pair by examining verb semantics for the presence of synonymy, near-synonymy, negation or antonymy in the subsequence.

In addition to these measures, there is also a *task* feature which identifies the task definition from which the sentence pair was derived. This allows the system to build separate classifiers for each task which we hoped would capture the different aspects of entailment specific to each task.

We investigated the usefulness of a number of distinct features during the development of our decision tree approach to textual entailment. These features were developed using the training part of the corpus made available for the PASCAL Recognising Textual Entailment Workshop (Dagan et al. 2005).[1] We describe the corpus in Sect. 4. Not all of these features were contributing factors in our final classification systems, but we list all of them here for the sake of completeness because some features are combinations of other atomic features. Table 1 gives a list of the features we used.

The first of our equivalence features are derived using the *ROUGE metrics*, which were used as a means of evaluating summary quality against a set of human-generated summaries in the 2004 Document Understanding Conference workshop (Duc 2004). The metrics provide a measure of word overlap (i.e. unigram, bigram, trigram and 4-gram), and a weighted and unweighted longest common subsequence measure.

The next semantic equivalence feature is calculated using the *cosine similarity measure*, which calculates the distance (or cosine of the angle) between the text/hypothesis pair in an n-dimensional vector space.

Using a *Latent Semantic Indexing* matrix constructed using the DUC 2004 corpus, we attempted to identify words in entailment pairs which have high co-occurrence statistics. This is an enhancement of the similarity measure given by the WordNet features, as it not

[1] The corpus may be downloaded from: http://www.pascal-network.org/Challenges/RTE/Datasets

only matches synonymy in the plain text, but also uses data from other corpora to identify other latent relationships.

WordNet was used to identify entailment between sentence pairs where corresponding synonyms are used. Words from the same synset (set of one or more synonyms, as defined by WordNet) were considered to indicate a greater likelihood of entailment. We believe that the accuracy of this feature could be greatly improved by disambiguating the sentence pair before calculating synset overlap. More specifically, in some instances multiple senses of a single term could be matched with terms in the corresponding entailment pair, resulting in sentences appearing more semantically similar than they actually are.

VerbOcean is a lexical resource that provides fine-grained semantic relationships between verbs. These related verb-pairs were gleaned from the web using lexical-syntactic patterns that captured five distinct verb relationships:

- similar-to (e.g. *escape*, *flee*)
- strength (e.g. *kill* is stronger than *wound*)
- antonymy (e.g. *win*, *lose*)
- enablement (e.g. *fight*, *win*)
- happens-before (e.g. *marry* happens before *divorce*)

VerbOcean also lists relationship strengths between verb pairs. In our experiments we only use the antonym and similar-to relationships for verb semantics analysis.

We also identify *adverbial negation* in the sentences. Adverbial negation occurs where the presence of a word (e.g. "nor", "not") modifies the meaning of the verb in the sentence. The information is gathered by simply counting the occurences of certain words in the text. We generate three features from this information:

- *negation_t* counts the number of occurrences of adverbial negation in the text
- *negation_h* counts the number of occurrences of adverbial negation in the hypothesis
- *negdiff* is the difference between negation_t and negation_h

Examination of the development set suggested that for a significant proportion of sentence pairs, the *longest common subsequence*[2] is largely similar to the hypothesis element, i.e. most of the hypothesis is contained in the text element. For this feature, we only examined verb semantics in the substrings that contain the longest common subsequence of the two sentences rather than in the full sentences. An example is shown in Fig. 1. There are three variations of this feature: lcs, lcs_pos and lcs_neg.

- *lcs* This feature holds one of three values $\{-1, 0, 1\}$, which correspond to the presence of an antonym, no relationship, or a synonym relationship between the longest common subsequence of the text and the hypothesis sentence, respectively
- *lcs_pos* and *lcs_neg* are simpler features which indicate the presence of a synonym relationship or antonym relationship, respectively

Using part-of-speech information, we identify the verbs in the text and hypothesis substrings. Direct comparison of these verbs gives the scores for lcs_pos and lcs_neg and, in combination, lcs. $lcs + not$ is another feature based on the longest common subsequence. It combines the above lcs features and also looks for the presence of words such as "not", which reverse the meaning of the sentence. Thus, for example, if an antonym and "not" occur in a sentence then this is considered to be a positive indication of entailment. Even though lcs+not is a combination of our lcs features we still retain the simpler features as it has been found that they improve entailment accuracy.

[2] The Longest Common Subsequence of a sentence pair is the longest (not necessarily contiguous) sequence of words which is common to both text and hypothesis.

id=1954; task=PP; judgement=FALSE
Text: *France on Saturday* flew *a planeload of United Nations aid into eastern Chad* where French soldiers prepared to deploy from their base in Abeche towards the border with Sudan's Darfur region.
Hypothesis:*France on Saturday* crashed *a planeload of United Nations aid into eastern Chad*

Fig. 1 Longest common subsequence. Italics denote the longest common subsequence

4 Evaluation

4.1 Corpora

We use three corpora for our experiments. The RTE corpus was developed for the Pascal RTE Challenge (Dagan et al. 2005). The corpus consists of three parts: two development sets which were released for training purposes during the system development stage, and one large test set used for evaluation of the participating systems. In each set positive and negative examples were divided into a number of different NLP tasks where textual entailment is used. These tasks include:

- CD: Comparable Documents
- IE: Information Extraction
- IR: Information Retrieval
- MT: Machine Translation
- QA: Question Answering
- RC: Reading Comprehension
- PP: Paraphrasing

The two development sets contained 287 and 280 sentence-pairs, respectively. The evaluation set contained 800 sentence-pairs. Each sentence-pair consists of a text and a hypothesis. The sentences come from datasets and corpora pertaining to the different NLP tasks. (see Figs. 1, 3–6 for examples of the sentence-pairs in the corpus.)

The second corpus we use was developed by Microsoft Research (Dolan et al. 2005). It consists of training and test sets which contain 4,076 and 1,725 pairs, respectively. The sentences were taken from online news articles which had been clustered by topic.

Manual investigation of this corpus showed a bias of approximately 2:1 in favour of positive classification. We have found that any bias in the training data tends to be reflected in the classifications produced by our system, e.g. the more negative examples in the training data, the more likely it is that the decision tree will return a negative classification. To investigate this further, we built a third corpus by removing positive instances from the MSR corpus until there were an equal number of positive and negative instances. Our modified corpus (MSR-5050) contains 2,646 pairs.

4.2 Evaluation metrics

We use a simple measure of accuracy to rate system performance. We examine the output classification of our systems, and determine the number of True and False classifications and how many of each were correct. We then define *accuracy* to be the number of correct classifications as a percentage of the total number of instances.

4.3 Experimental methodology

We wished to show that our current system, which focuses on a number of different aspects of sentence-pairs, is more effective than our previous system which used only the cosine and LSI features (Newman et al. 2004). Therefore, we have two classifiers which model these systems. Both classifiers are trained and tested on each of the three corpora described above.

4.4 Results

Examining the accuracy results for each corpus (see Table 2), we see that there is an improvement in performance on each corpus when the additional syntactic and lexical features are used.

On the RTE corpus, we achieved an improvement in performance of almost 22% over our baseline entailment system (see Fig. 2). In contrast, there are relatively modest increases in performance for the MSR and MSR-5050 corpora. The MSR-5050 corpus showed a larger rise in performance (4%) than the 2:1-biased MSR corpus (1%). Since MSR-5050 is a subset of the latter, this supports our assertion (see end of Sect. 4.1) that the bias of the training set is an important factor when using a decision tree classifier for this task. Since the features we currently use were developed using a balanced dataset as reference, this is reflected in the different rates of improvement.

Empirical investigation shows that certain cohorts (e.g. Comparable Documents) of the RTE corpus are relatively simple to classify, based on certain features such as word-overlap

Table 2 Classifier accuracy on each corpus, using original and new features

Corpus	Original system (%)	New system (%)
RTE	52.25	74.00
MSR	67.71	68.75
MSR-5050	59.71	63.94

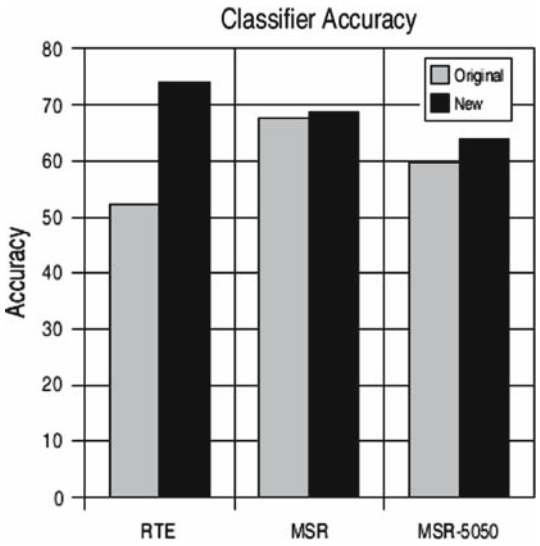

Fig. 2 Comparison of classifier accuracy of each system on all three corpora

id=1560; task=QA; judgement=TRUE
Text: The technological triumph known as GPS – the Global Positioning System of satellite-based navigation – was incubated in the mind of Ivan Getting.
Hypothesis: Ivan Getting invented the GPS.

id=858; task=CD; judgement=TRUE
Text: Each hour spent in a car was associated with a 6 percent increase in the likelihood of obesity and each half-mile walked per day reduced those odds by nearly 5 percent, the researchers found.
Hypothesis: The more driving you do means you're going to weigh more – the more walking means you're going to weigh less.

Fig. 3 Compositional paraphrases (misclassified by our system)

id=2028; task=QA; judgement=FALSE
Text: *Besancon is the capital of France*'s watch and clock-making industry and of high precision engineering.
Hypothesis: *Besancon is the capital of France.*

id=1964; task=PP; judgement=FALSE
Text: Under the avalanche of Italian outrage *London Underground* has apologised and agreed *to withdraw the poster.*
Hypothesis: *London Underground* opposed *to withdraw the poster.*

Fig. 4 LCS features

id=828; task=CD
Text: Jennifer Hawkins is the 21-year-old beauty queen from Australia.
Hypothesis: Jennifer Hawkins is Australia's 20-year-old beauty queen.

id=868; task=CD
Text: Several other people, including a woman and two children, suffered injuries in the incident.
Hypothesis: Several people were slightly wounded, including a woman and three children.

Fig. 5 Numerical examples (misclassified by our system)

id=1432; task=PP
Text: *The report* faults *intelligence agencies* for a "lack of imagination" in not anticipating that al Qaeda could attack the United States using hijacked aircraft.
Hypothesis: *The report* backs *intelligence agencies.*

Fig. 6 LCS + not feature in action

and the presence of negation. Further investigation shows that those sentences with highest cosine similarity (i.e. greatest number of shared words) were, rather counter-intuitively, more likely to be cases of *negative* entailment rather than positive entailment. Therefore, our features designed to recognise the the presence of negation indicators (e.g. "not", antonymy) are highly informative on the RTE corpus. Their introduction brought about an improvement in performance from 52% accuracy to 74% accuracy over our old system which classified on word similarity alone.

By contrast, the MSR corpus, having been drawn from real-world news sources, contains a much more realistic distribution of instances (i.e. those with high word overlap are likely to have high semantic equivalence). Thus we see little difference in performance between the two systems on this corpus, because the most informative features are the word-overlap metrics, which are common to both systems.

5 Discussion of classification errors

In this section we discuss, with examples, some common system errors made by our decision tree classifier.

5.1 Compositional paraphrases and syntactic paraphrases

Recognition of textual entailment is difficult in the case where sentence structure has been changed (thus nullifying our lcs metrics), or where synonyms are extensively used (where we must rely on WordNet and VerbOcean to identify if a relationship exists). It is clear from our system description in Sect. 3 that the majority of our features deal with the identification of word-level, atomic paraphrase units (e.g. child = kid; eat = devour). Consequently, there are a number of examples where phrasal and compositional paraphrasing has resulted in misclassifications by our system. Some examples of this are shown in Fig. 3.

Another important type of paraphrase, not addressed explicitly by our system, is the syntactic paraphrase (e.g. "I ate the cake" or "the cake was eaten by me"). However, although we didn't include a parse tree analysis in our approach, it appears that the ROUGE metrics (and to some extent the cosine metric) were an adequate means of detecting syntactic paraphrases. The position of the ROUGE features in high-level nodes in the decision tree confirms that n-gram overlap is an important aspect of textual entailment, but obviously not enough on its own.

We also observed that in some cases syntactic paraphrases prevented the detection of longest common subsequences, and reduced the effectiveness of features that relied on this syntactic analysis. Consequently, parse tree analysis and subsequent normalisation of sentence structure could be an effective solution to this problem.

5.2 Modifying pre-texts and post-texts

Our LCS-based features focus, by definition, on the parts of the text and hypothesis which are most common to both.

Overall, our LCS-based features were critical to the classification decision; however, we did find instances where sentence pairs were misclassified by over-simplification of the textual entailment task. For example, pair 2,028 in Fig. 4 shows how the true meaning of the text sentence can extend beyond the longest common subsequence. In addition, pair 1,964 shows how coverage limitations in the VerbOcean resource resulted in this example being misclassified as negative, because an antonym relationship between "agree" and "oppose" was not listed.

5.3 Numerical strings

During our manual examination of the results we also noticed another crucial analysis component missing from our system: numerical string evaluation. Two examples are shown in

Fig. 5. Future development will focus on a normalisation method for evaluating numeric values in the entailment pair.

5.4 The effectiveness of lcs + not feature

The example in Fig. 6 illustrates the effectiveness of our lcs + not feature overriding the generally more dominant lcs score. As we can see, there is a strong overlap between the text and hypothesis. However the presence of the antonym pair in the middle of the subsequences acts to negate any entailment, and our system returns a result of "false" in these cases.

Unfortunately, there are a number of similar cases to this in the corpus which we classified incorrectly. This may be due to the coverage limitations of the knowledge repositories we are using (i.e. WordNet, VerbOcean) to recognise all correct synonym and antonym verb relations.

6 Conclusion

In this paper, we showed how Textual Entailment Recognition can contribute to various areas of research in Natural Language Processing. We showed that the syntactic and semantic features provide an improvement in system performance compared to the purely lexical features. This was shown to be especially true for sentence-pairs derived from such tasks as Comparable Documents (CD) and Paraphrase Acquisition (PP).

We have shown that the features added to our system do bring about an improved performance. However this improvement is very small for the MSR corpora, and indeed, for a number of cohorts of the RTE corpus. From this we conclude that while the work done so far has provided an incremental advance, any real improvement will be the result of significant breakthrough in how the problem is modelled. Future work will be directed towards the improvement and refinement of current features, the development of new features to cover aspects of our current model and we must continue to investigate the domain for new clues to find a better path to a solution.

We also plan to evaluate our system by judging its performance in two challenging NLP applications: question–answering and multi-document summarisation. Our aim will be to show that the identification of semantically-equivalent sentences using our technique can improve the overall performance of our system on these tasks. The work for these applications will be done as part of the CASIA project, which aims to harness NLP techniques in the field of Incident and Accident Analysis.

Acknowledgement The authors gratefully acknowledge the support of Enterprise Ireland in their support of this research as part of *CASIA – Combined Approaches to Summarisation for Incident Analysis* (Project Number SC/2003/0255).

References

Akhmatova E (2005) Textual entailment resolution via atomic propositions. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK

Bos J, Markert K (2005) Combining shallow and deep NLP methods for recognizing textual entailment. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK

Budanitsky A, Hirst G (2001) Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In: Proceedings of workshop on wordNet and other lexical resources, second meeting of the North American chapter of the association for computational linguistics, Pittsburgh, PA, USA

Chklovski T, Pantel P (2004) VerbOcean: mining the web for fine-grained semantic verb relations. Proceedings of conference empirical methods in natural language processing (EMNLP-04), Barcelona, Spain

Corley C, Mihalcea R (2005) Measuring the semantic similarity of texts. In: Proceedings of ACL workshop on empirical modelling of semantic equivalence and entailment, ACL, Ann Arbor, Michigan, USA

Dagan I, Glickman O, Magnini B (eds) (2005) Proceedings of the PASCAL challenges workshop on recognising textual entailment

Deerwester S, Dumais S, Furna G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inform Sci 41(6):391–407

Dolan W, Dagan I (eds) (2005) Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment. June 30th 2005, Ann Arbor, Michigan, USA

Dolan W, Brockett C, Quirke C (2005) Microsoft research paraphrase corpus. At http://research.microsoft.com/~billdol/

Document Understanding Conference (DUC) National Institute of Standards and Technology, USA. At http://duc.nist.gov.

Fellbaum C (ed) (1998) WordNet: an electronic lexical database. Available from MIT Press

Fowler A, Hauser B, Hodges D, Niles I, Novischi A, Stephan J (2005) Applying COGEX to recognize textual entailment. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK

Herrera J, Peñas A, Verdejo F (2005) Textual entailment recognition based on dependency analysis and WordNet. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK

Lin C-Y, Hovy E (2004) Automatic evaluation of summaries using n-gram co-occurence statistics. Proc. document understanding conference (DUC). National Institute of Standards and Technology

Marsi E, Krahmer E (2005) Classification of semantic relations by humans and machines. In: Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment, Southampton, UK

Newman E, Stokes N, Carthy J, Dunnion J (2004) Comparing redundancy removal techniques for multi-document summarisation. In: Proceedings of STAIRS 2004, Valencia, Spain

Pazienza MT, Pennacchiotti M, Zanzotto FM (2005) Textual entailment as syntactic graph distance. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK

Quinlan JR (2002) C5.0 machine learning software. At http://www.rulequest.com

Raina R, Haghighi A, Cox C, Finkel J, Michels J, Toutanova K, MacCartney B, de Marneffe C-M, Manning CD, Ng A (2005a) Robust textual inference using diverse knowledge sources. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK

Raina R, Ng A, Manning CD (2005b) Robust textual inference via learning and abductive reasoning. American Association of Artificial Intelligence

de Salvo Braz R, Girju R, Punyakanok V, Roth D, Sammons M (2005) An inference model for semantic entailment in natural language. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK

Vanderwende L, Coughlin D, Dolan B (2005) What syntax can contribute in entailment task. In: Proceedings of the PASCAL challenges workshop on recognising textual entailment, Southampton, UK