The Use of Sentence Similarity as a Semantic Relevance Metric for Question Answering

Marco De Boni
Department of Computer Science, University of York, York Y010 5DD, United Kingdom
and School of Computing, Leeds Metropolitan University, Leeds LS6 3QS, United Kingdom
mdeboni@cs.york.ac.uk, m.deboni@lmu.ac.uk

Suresh Manandhar
Department of Computer Science, University of York, York Y010 5DD, United Kingdom
suresh@cs.york.ac.uk

Abstract

An algorithm for calculating semantic similarity between sentences using a variety of linguistic information is presented and applied to the problem of Question Answering. This semantic similarity measure is used in order to determine the semantic relevance of an answer in respect to a question. The algorithm is evaluated against the TREC Question Answering test-bed and is shown to be useful in determining possible answers to a question. Not all linguistic information is shown to be useful, however, and an in-depth analysis shows that certain sentence features are more important than others in determining relevance.

Semantic relevance for Question Answering

Question Answering Systems aim to determine an answer to a question by searching for a response in a collection of documents (see Voorhees 2002 for an overview of current systems). In order to achieve this (see for example Harabagiu et al. 2002), systems narrow down the search by using information retrieval techniques to select a subset of documents, or paragraphs within documents, containing keywords from the question and a concept which corresponds to the correct question type (e.g. a question starting with the word "Who?" would require an answer containing a person). The exact answer sentence is then sought by either attempting to unify the answer semantically with the question, through some kind of logical transformation (e.g. Moldovan and Rus, 2001) or by some form of pattern matching (e.g. Soubbotin 2002; Harabagiu et al. 2002).

Semantic relevance (Berg 1991) is the idea that an answer can be deemed to be relevant in respect to a question in virtue of the meanings of the words in the answer and question sentences. So, for example, given the question

Q: What are we having for dinner?

And the possible answers

A1: Spaghetti

A2: Tomorrow.

A1 would be considered more semantically relevant because the meaning of the word dinner is more related to the word "spaghetti" (one eats at dinner, and spaghetti is something to eat) than to the word "tomorrow". Here we investigate the use of semantic relevance for the problem of Question Answering, by widening the context considered to judge relevance to the whole sentence which contains an answer in a collection of documents.

We therefore argue that Question Answering Systems therefore make implicit use of the notion of semantic relevance, the idea that a response can be deemed relevant in virtue of the meanings of the expressions in the sentence, in order to determine sentences which contain an answer to a given question. In particular, we propose a semantic relevance metric which does not rely on logical inference or unification and does not require the extensive work required for handcrafted pattern matching rules and is instead an extension of the idea of paragraph similarity.

Previous Work

To talk about semantic relevance is to talk about the semantic similarity between sentences. WordNet (Miller 1995; Fellbaum 1998), a lexical database which organizes words into synsets, sets of synonymous words, and specifies a number of relationships such as hypernym, synonym, meronym which can exist between the synsets in the lexicon, has been shown to be fruitful in the calculation of semantic similarity. One approach has been to determine similarity by calculating the length of the path or relations connecting the words which constitute sentences (see for example Green 1997); different approaches have been proposed (for an evaluation see Budanitsky and Hirst 2001), either using all WordNet relations (Hirst-St-Onge 1998) or only is-a relations (Resnik 1995; Jiang-Conrath 1997; Lin 1998). Mihalcea and Moldovan (1999), Harabagiu et al. (1999) and De Boni and Manandhar (2002) found WordNet glosses, considered as micro-contexts, to be useful in determining conceptual similarity. Lee et al. (2002) have applied conceptual similarity to the Question Answering task, giving an answer A a score dependent on the number of matching terms in A and the question.

Here we follow on these ideas, evaluating different approaches to the measure of semantic relevance for question answering, showing how the use of WordNet information, part-of-speech information and head-phrase chunking can provide a reliable measurement for semantic relevance.

A sentence similarity metric for Question Answering

In order to determine the relevance of a possible answer to a question, a basic semantic relevance algorithm was employed to calculate the semantic similarity between the question sentence and the answer sentence. This was then augmented to evaluate what additions could improve the measure.

Basic Semantic Relevance Algorithm

Semantic similarity was considered as a function which took as arguments a question q and an answer a and returned a value representing the semantic relevance of a in respect of q, i.e.

$$\text{semantic-relevance}(a, q) = n \in \mathbb{R}$$

The co-domain *semantic-relevance*(a, q) is therefore made up of the real numbers $\mathbb{R}$, interpreted as: *semantic-relevance*(a_1,q) < *semantic-relevance*(a_2,q) represents the fact that answer a_1 is less relevant than a_2 in respect to the question q and the context B.

In order to calculate the semantic similarity between a question q and an answer q, q and a were considered as sets A and Q of word stems. The similarity between each word in the question and each word in the answer was then calculated and the sum of the closest matches gave the overall similarity. In other words, given two sets Q and A, where

$$Q=\{qw_1,qw_2,\ldots,qw_n\} \text{ and}$$

$$A=\{aw_1,aw_2,\ldots,aw_m\},$$

the similarity between Q and A is given by

$$\sum_{1<p<n} \text{Argmax}_m \text{ similarity}(qw_p, aw_m)$$

The function *similarity*(w_1, w_2) maps the stems of the two words w_1 and w_2 to a similarity measure m representing how semantically related the two words are. The co-domain *similarity*(w_1, w_2)is made up of the real numbers $\mathbb{R}$ in the interval 0..1; *similarity*(w_i, w_j)< *similarity*(w_i, w_k) represents the fact that the word w_j is less semantically related than w_k in respect to the word w_i. In particular *similarity*=0 if two words are not at all semantically related and *similarity*=1 if the words are the same.

$$similarity(w_1, w_2) = h \in \mathbb{R}$$

where

$$0 \leq h \leq 1$$

The *similarity* function is calculated as follows:

$$similarity(w_1, w_2) = 0 \text{ if } w_1 \in ST \vee w_2 \in ST$$

where ST is a set containing a number of stop-words (e.g. “the”, “a”, “to”) which are too common to be able to be usefully employed to estimate semantic similarity.

Else

$$similarity(w_1, w_2) = s$$

where s is calculated as follows: the words w_1 and w_2 are then compared using all the available WordNet relationships (is-a, satellite, similar, pertains, meronym, entails, etc.), with the additional relationship, “same-as”, which indicated that two words were identical. Each relationship is given a weighting indicating how related two words are, with a “same as” relationship indicating the closest relationship, followed by synonym relationships, hypernym, hyponym, then satellite, meronym, pertains, entails.

So, for example, given the question:

“Who went to the mountains yesterday?”

and the possible answer

“Fred walked to the big mountain and then to mount Pleasant”

Q would be

{who, go, to, the, mountain, yesterday}

A would be

{Fred, walk, to, the, big, mountain, and, then, to, mount, Pleasant}

In order to calculate similarity the algorithm would consider each word in turn. “Who” would be ignored as it is a common word and hence part of the list of stop-words. “Go” would be related to “walk” in a is-a relationship and receive a score s_1. “To” and “the” would be found in the list of stop-words and ignored. “Mountain” would be considered most similar to “mountain” (same-as relationship) and receive a score s_2: “mount” would be in a synonym relationship with “mountain” and give a lower score, so it is ignored. “Yesterday” would receive a score of 0 as there are no semantically related words in Q. The similarity measure of Q in respect to A would therefore be given by $s_1 + s_2$.

Enhanced Algorithm

The basic algorithm was augmented to consider the following information about sentences:

NP chunking. The intuition behind the use of NP chunks to determine similarity is the fact that the phrase “When did [the big bad wolf] eat [red riding hood] ?” is more similar to “[the bad wolf] ate [red riding hood] [who] carried [a pink chicken]” than “[the wolf] who wore [a pink riding hood] ate [the bad red chicken]”. Words appearing in similar NPs in both question and answer were therefore considered more relevant than words that did not.

Part-of-Speech tagging. The intuition behind the use of Part-of-Speech tagging is that this should disambiguate the senses of the words sufficiently to avoid gross errors in determining similarity, for example when considering the word relative in the two phrases “the facts relative [adjective] to the issue” and “John visited his relatives [noun]”.

Compound noun information. The motivation behind is similar to the reason for using chunking information, i.e. the fact that the word “United” in “United States” should not be considered similar to “United” as in “Manchester United”. As opposed to when using chunking information, however, when using noun compound information, the compound is considered a single word, as opposed to a group of words: chunking and compound noun information may therefore be combined as in “[the [United States] official team]”.

Proper noun information. The intuition behind this is that titles (of books, films, etc.) should not be confused with the “normal” use of the same words: “blue lagoon” as in the sentence "the film Blue Lagoon was rather strange" should not be considered as similar to the same words in the sentence "they swan in the blue lagoon" as they are to the sentence "I enjoyed Blue Lagoon when I was younger".

Word frequency information. This is a step beyond the use of stop-words, following the intuition that the more a word is common the less it is useful in determining similarity between sentence. So, given the sentences "metatheoretical reasoning is common in philosophy" and "metatheoretical arguments are common in philosophy", the word "metatheoretical" should be considered more important in determining relevance than the words "common", "philosophy" and "is" as it is much more rare and therefore less probably found in irrelevant sentences.

Evaluation

Method

The semantic relevance measure was used to rank answer sentences for a total of 200 questions randomly chosen from the test questions used in the NIST TREC-10 Question Answering track (Voorhees 2002). The answers sentences were sought in the set of 50 documents also provided by NIST for each questions, which consisted of newspaper articles from the Los Angeles Times, the New York Times, the Financial Times, AP newswire and the San Jose Mercury; which should have (but were not guaranteed to) contained an answer to the questions. The documents were manually checked in order to verify that an answer was actually present in the texts and questions for which no answer was given in the documents were discarded (about 32% of the questions did not have an answer in the given set of documents). The sentences were then ranked using the above the semantic relevance algorithm and the first 15 answer sentences were manually checked to establish at which rank (if any) an answer sentence was found.

A correct answer sentence was strictly taken to be a sentence which "justified" an answer, i.e. from which an intelligent reader could infer an answer without the use of any knowledge which could be employed to derive an answer without the need to consult the answer document. This meant that in a number of instances sentences which could have been deemed to contain an answer were judged as non-answers as they did not justify the answer. So, for example, given the question "When was the first stamp issued", the sentence "A block of "penny black" postage stamps, issued in 1840, were sold at auction Tuesday for about $442,680, an auction house said." was not judged to contain an answer, as the correct answer (1840) could not be inferred from this sentence alone.

Results

A number of runs were carried out in order to determine what features would best determine relevance (results are summarised in Figure 1). The features examined were:

- NP bracketing. This could be either:
  - Present in full ("chunked")
  - Absent ("dumb")
  - Naïve (see below for an explanation)
- Part of Speech Tagging, which could be either
  - Present
  - Absent
  - Flexible

A series of experiments were then carried out with "normal" part-of-speech tagging but with the addition of the following features:

- Word frequency information
- Proper noun and compound word information.

For each set of features the ranked answers were examined manually to identify the exact rank of sentence containing an answer. In line with the type of evaluation carried out for the TREC QA workshop (Voorhees 2002), the mean reciprocal rank (MRR), i.e. the average reciprocal of the rank at which the answer was found, was then computed for the given results.

Rank	Basic algorithm with POS			Basic algorithm , no POS			Basic algorithm , flexible POS		
	chunked	"dumb"	Naïve	chunked	"dumb"	Naïve	chunked	“dumb”	Naïve
Ans. 1 (%)	18.4	28.6	24.5	14.3	24.5	16.3	28.6	24.5	24.5
Ans. 1-5 (%)	51	55.1	55.1	55.1	53	51	57.1	53	49
MRR	0.436	0.505	0.465	0.433	0.52	0.43	0.519	0.347	0.496

Rank	Basic algorithm , with POS and word frequency		Basic algorithm , with POS, word frequency, compounding and proper noun identification		Simple Word Match
	chunked	“dumb”	chuncked	“dumb”	
Ans. 1 (%)	20.4	28.6	24.5	30.7	18
Ans. 1-5 (%)	51	53	53	57.1	37
MRR	0.434	0.451	0.481	0.484	0.2

Fig. 1

Compared to a simple word match, the semantic relevance algorithm is a great improvement. However the individual features contribute in different measure to this result. Initial experiments indicated that NP chunking did not contribute to improving performance. It was speculated that this was in large part due to the POS tagger, which incorrectly tagged adjectives and nouns in a number of occasions. In order to verify this, experiments were carried out with “flexible” tagging which allowed words to be considered both as nouns and adjectives. With this correction NP chunking gave the best results. Another possible cause for the poor results in NP chunking was incorrect chunking. In order to verify this, experiments were carried out using a simple NP chunker that gathered together words found between articles and verbs. This gave better results when using the standard POS tagging, but worse results in all other cases, indicating that the given NP chunker was indeed better that a simpler chunker. Further analysis of the questions indicated that the reason NP chunking did not prove as useful as initially hoped was the type of questions given, which were usually very short, sometimes with NPs consisting of no more than one word.

The addition of information about Proper nouns (taken to be words beginning with a capital letter), compound nouns (taken to be compounds identified by WordNet, e.g. “United States”) significantly improved performance. It is important to note that recognising compound nouns gave different information from the recognition of NP chunks: a simple example is the NP “the Coca Cola company” where the compound noun was taken to be simply “Coca Cola”.

Experiments were also carried out on the use of word frequency information. Given that the questions examined were generic queries which did not necessarily refer to a specific set of documents, the word frequency for individual words was taken to be the word frequency given in the British National Corpus (see BNCFreq 2002 for a list of the most frequent 3000 words in English according to the British National Corpus). The top 100 words (see BNCFreq 2002), making up 43% of the English Language, were then used as stop-words and were not used in calculating semantic similarity. The similarity between words was therefore weighted to take into account the commonality of the words in generic English. This approach also gave good results. A question answering system tailored for a specific task, would however have to use a word frequency table specific to that task: "analgesic" may be an infrequent term in everyday usage, but is quite common in medical documents. While word frequency improved the accuracy of the mesurement not employing NP chunking, this approach had a detrimental effect when combined with NP chunking, probably due to the fact that "chunks" containing very common words were not able to furnish any useful information gain.

Finally, the addition of information about compound words (e.g. "compound interest") and information about which words proper nouns provided the best results.

Discussion

When looking for a definite answer, semantic relevance appears to be helpful: the best results indicated that a sentence containing an answer would ranked as the top answer sentence in the document collection in 30.7% of cases. An answer sentence would be ranked amongst the top five most relevant sentences in 57.1% of cases. In most of the remaining cases (around 90%) an answer was found in the top fifteen sentences considered relevant, but the ranking could not be correct due to very short questions (e.g. "Where is Perth?"): indeed, in a number of cases all sentences were assigned the same similarity score. In other cases the answer sentence required additional background knowledge that was not provided by WordNet: examples which caused answers to be missed are the lack of any connection between the words "midway" and "middle" and the very tenuous connection between "man made object" and satellite, which requires an elaborate traversal of a number of different semantic relations. In other cases some form of non-trivial reasoning would have been necessary to deduce an answer: for example, there was no answer in the set of documents to the question "During which season do most thunderstorms occur?", but an intelligent reader could have counted the number of times thunderstorms were mentioned in articles about the weather and deduced a statistical analysis of the occurrence of thunderstorms during the year. Yet more questions required assumptions about questioners' intentions, e.g. the question "What is Wimbledon?", which could have had as answer "A part of London", or "A tennis competition".

To get an idea of the performance that a system which used this algorithm could achieve, the best system at the TREC 2002 QA workshop (Soubbotin 2002), which employed a very rich and complex system of pattern matching, found an answer in the top 5 sentences 69.1% of the time, with a MRR of 0.676; the top five systems had an average MRR of 0.518 and an answer in the top five sentences on average 62.1% of the time. Simply using documents provided by a standard Information Retrieval engine, the semantic relevance algorithm presented correctly ranked the sentence containing an answer in the top 5 sentences 57.1% of the time, with a MRR of 0.519. Combining this semantic relevance algorithm with the techniques used by the TREC participants (complex pattern matching, precise question-type and named-entity recognition, simple forms of logical inference or abduction) could therefore give some very positive results.

Conclusions and Further work

Semantic relevance therefore does provide a useful indication to which sentences in a document collection contain the answer to a question. On its own, however, this metric cannot provide an answer in all cases and other techniques such as the additional of domain-specific knowledge, logical inference and information about user goals must be used to determine the actual answer in the remaining cases. Future work will examine the integration of the given semantic relevance metric with additional background knowledge and the use of inferencing and knowledge about user goals to deduct answers.

Bibliography

BNCFreq. 2002. (last accessed June 2002). http://www.eecs.umich.edu/~qstout/586/bncfreq.html.

Budanitsky, A., and Hirst, G. 2001. "Semantic distance in WordNet: and experimental, application-oriented evaluation of five measures", in Proceedings of the NAACL 2001 Workshop on WordNet and other lexical resources, Pittsburgh.

De Boni, M. and Manandhar, S., 2002. "Automated Discovery of Telic Relations for WordNet", Proceedings of the First International WordNet Conference, India.

Fellbaum, C., 1998. "WordNet, An electronic Lexical Database", MIT Press.

Green, S. J., 1997. "Automatically generating hypertext by computing semantic similarity", Technical Report n. 366, University of Toronto.

Harabagiu, S. A., Miller, A. G., Moldovan, D. I., 1999. "WordNet2 - a morphologically and semantically enhanced resource", In Proceedings of SIGLEX-99, University of Maryland.

Harabagiu, S., et al., 2002. "Answering Complex, List and Context Questions with LCC's Question-Answering Server", Proceedings of TREC-10, NIST.

Hirst, G., and St-Onge, D. 1998. "Lexical chains as representations of context for the detection and correction of malapropisms", in Fellbaum (ed.), "WordNet: and electronic lexical database", MIT Press.

Jiang, J. J., and Conrath, D. W. 1997. "Semantic similarity based on corpus statistics and lexical taxonomy", in Proceedings of ICRCL, Taiwan.

Lee, G. G., et al., 2002. "SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP", Proceedings of TREC-10, NIST,

Lin, D., 1998. "An information-theoretic definition of similarity", in Proceedings of the 15th International Conference on Machine Learning, Madison.

Mihalcea, R. and Moldovan, D. 1999. "A Method for Word Sense Disambiguation of Unrestricted Text", in Proceedings of ACL '99, Maryland, NY.

Miller, G. A. 1999. "WordNet: A Lexical Database", Communications of the ACM, 38 (11).

Moldovan, D. and Rus, V. 2001. "Logic Form Transformation of WordNet and its Applicability to Question Answering", Proceedings of the 39th conference of ACL, Toulouse.

Resnik, P. 1995. "Using information content to evaluate semantic similarity", in Proceedings of the 14th IJCAI, Montreal.

Soubbotin, M.M., 2002. "Patterns of Potential Answer Expressions as Clues to the Right Answers", Proceedings of TREC-10, NIST.

Voorhees, E. 2002. "Overview of the TREC 2001 Question Answering Track", Proceedings of TREC-10, NIST.