An Entailment-Based Approach to the QA4MRE Challenge

Peter Clark[1], Phil Harrison[1], Xuchen Yao[2]

[1] Vulcan Inc, Seattle, WA 98104 ({peterc,philipha}@vulcan.com)
[2] Johns Hopkins Univ, Baltimore, Maryland 21218 (xuchen@cs.jhu.edu)

Abstract. This paper describes our entry to the 2012 QA4MRE Main Task (English dataset). The QA4MRE task poses a significant challenge as the expression of knowledge in the question and answer (in the document) typically substantially differs. Ultimately, one would need a system that can perform full machine reading – creating an internal model of the document's meaning – to achieve high performance. Our approach is a preliminary step toward this, based on estimating the likelihood of textual entailment between sentences in the text, and the question Q and each candidate answer A_i. We first treat the question Q and each answer A_i independently, and find sets of sentences SQ, SA that each plausibly entail (the target of) Q or one of the A_i respectively. We then search for the closest (in the document) pair of sentences $<S_Q \in SQ, S_{Ai} \in SA>$ in these sets, and conclude that the answer A_i entailed by S_{Ai} in the closest pair is the answer. This approach assumes coherent discourse, i.e., that sentences close together are usually "talking about the same thing", and thus conveying a single idea (namely an expression of the $Q+A_i$ pair).

In QA4MRE it is hard to "prove" entailment, as a candidate answer A may be expressed using a substantially different wording in the document, over multiple sentences, and only partially (as some aspects of A may be left implicit in the document, to be filled in by the reader). As a result, we instead estimate the likelihood of entailment (that a sentence S entails A) by look for evidence, namely entailment relationships between components of S and A such as words, bigrams, trigrams, and parse fragments. To identify these possible entailment relationships we use three knowledge resources, namely WordNet, ParaPara (a large paraphrase database from Johns Hopkins University), and the DIRT paraphrase database. Our best run scored 40% in the evaluation, and around 42% in additional (unsubmitted) runs afterwards. In ablation studies, we found that the majority of our score (approximately 38%) could be attributed to the basic algorithm, with the knowledge resources adding approximately 4% to this baseline score. Finally we critique our approach with respect to the broader goal of machine reading, and discuss what is needed to move closer to that goal.

1 Introduction

Machine Reading remains one of the Grand Challenges of Artificial Intelligence, and also one of the most difficult. Machine Reading requires more than just parsing a text; it requires constructing a coherent internal model of the world that the text is

adfa, p. 1, 2011.

describing. This goal is particularly challenging because typically only a fraction of that model is made explicit in text, requiring substantial background knowledge to fill in the gaps and resolve ambiguities (Schank and Abelson, 1977). By one estimate, only one eighth of the knowledge conveyed by text is stated explicitly (Graesser, 1981).

The QA4MRE task is a simpler version of the full Machine Reading challenge because it uses multiple-choice questions against a single document. However, QA4MRE is still formidable because the answer information is typically expressed in varied, imprecise, complex, and sometimes distributed ways in the document, and almost always requires background knowledge to bridge the gap to the original question and candidate answers. In the ideal case, a system would still build an internal model of the entire document, and decide which candidate answer that model entails. In our system, we use a simpler and more impoverished approach, namely to look for entailment relationships between phrases or sentences in the document and the question Q and candidate answers A_i, as a first step towards more complete model construction.

An overview of our approach is as follows: We first treat the question Q and each answer A_i independently, and find sets of sentences SQ, SA that each plausibly entail (the target of) Q or one of the A_i respectively. We then search for the closest (in the document) pair of sentences $<S_Q \in SQ, S_{Ai} \in SA>$ in these sets, and conclude that the answer A_i entailed by S_{Ai} in the closest pair is the answer. The justification for this is that sentences close together are usually "talking about the same thing", and thus conveying a single idea (namely an expression of the $Q+A_i$ pair).

In the rest of this paper we describe our approach, present some examples, and then summarize our results. Our best run scored 40% in the evaluation, and around 42% in additional (unsubmitted) runs afterwards. In ablation studies, we found that the majority of our score (approximately 38%) could be attributed to the basic algorithm, with the knowledge resources adding approximately 4% to this baseline score. Finally we critique our approach with respect to the broader goal of machine reading, and discuss what is needed to move closer to that goal.

2 Approach

2.1 Overview

One approach to QA4MRE would be to substitute a candidate answer A into a question Q to form a single sentence, called Q+A, and then assess whether the document semantically entails Q+A. However, in our early experiments with this approach, we had mediocre performance. We believe this was, in part, because this involves searching for several pieces of information at once (from within both Q and A), with some items of information confusing the search for others. As a result, we have found a different approach to be more effective:

1. find expressions of Q and A independently in the document
2. assess whether those two expressions indicate A is an answer to Q

By searching for Q and A independently in the document, we avoid answer details confusing the search for the question statement in the document, and vice versa.

These two steps rely on two important tasks:

Task 1. Assessing how likely it is that a sentence[1] is expressing the same information as (the target of) a question Q or a candidate answer A

Task 2. Assessing how likely it is that an expression of Q and an expression of A imply that A is an answer to Q.

Task 1 can be viewed as a form of the Textual Entailment challenge:

- Given sentence S and (the target of) a question Q, does S entail (the target of) Q?
- Given sentence S and a candidate answer A, does S entail A?

By "target of the question", we mean the description of the item the question is asking for, typically a noun phrase. For instance in "Where is the epicenter of the AIDS pandemic?" the target is "the epicenter of the AIDS pandemic".

For example, given the Q+A pair and the sentence S37 below from the target document:

Question Q[3.11] Why were transistor radios a significant development?
Answer A2. Young people could listen to pop outside.
Sentence S27. In the 1960s, the introduction of inexpensive, portable transistor radios meant that teenagers could listen to music outside of the home.

the entailment questions our current system asks are:

- How likely is that that S27 entails "transistor radios were a significant development" (Q)?
- How likely is it that S27 entails "Young people could listen to pop outside" (A2)?

Task 2 can also be viewed as a Textual Entailment task: Given sentence S1 plausibly entails Q, and sentence S2 plausibly entails A, does S1+S2 entail Q+A? (where Q+A is a sentence created by substituting A into Q). To assess this, we significantly approximate this task by simply measuring how close S1 and S2 are in the document, the proximity being taken as a measure of likelihood that S1+S2 entails Q+A. The justification for this is an assumption of coherent discourse, i.e., that sentences close together are usually "talking about the same thing", and thus close sentences are often conveying a single coherent idea (e.g., the Q+A pair). Although this is a gross approximation, it is helpful because often the connection between Q+A is not explicit in the document. Rather, it is implied by pragmatic considerations such as context, sentence ordering, or subtle discourse words (as in the above example).

Although Q and A2 in the above example are complete sentences, we apply the same approach when the Q and A are phrases (as is more usually the case). We say

[1] We assume that (the target of) a question or answer is expressed in a single sentence, although the expression of the two together may span multiple sentences.

that a sentence S "entails" a phrase P if the meaning of P forms part of the meaning of S. For example we say sentence S entails the answer phrase A below:

S (“Text”)**:** Because humanity has polluted so much surface water on the planet, we are now mining the groundwater far faster than it can be replaced by nature.
A (“Hypothesis”): Because surface water is too polluted.

because the notion that "surface water is too polluted" is part of the meaning of S.

We now describe Task 1 (Entailment Assessment) and Task 2 (Proximity Detection) in more detail.

2.2 Task 1: Entailment Assessment

To determine if a sentence S entails a candidate answer A, one approach is to create a formal representation of S and A and then prove S implies A. However, reliably creating formal representations is extremely difficult. A weaker but more practical variant is to do reasoning at the textual level - the "natural logic" approach (MacCartney & Manning, 2007; MacCartney, 2009) - in which semantically valid (or plausible) rewrite rules are applied directly to the linguistic structure of the text. If S's parse can be validly transformed to A's parse, then A is, in a way, "proved" (entailed). However, even this is a high bar; often we cannot fully "prove" that S entails A by this method, either because we are missing background knowledge, or because some unstated context/assumptions are needed, or because in a strict sense A is not fully derivable from S due to some of the required knowledge being implicit (unstated). As a result, we relax the problem further and collect *evidence* of entailment (do pieces of S entail pieces of A?) to provide an overall assessment of how likely S entails A. This is a standard approach taken in most Recognizing Textual Entailment (RTE) systems (e.g., NIST, 2011). The key question is which evidence should be used, and how that evidence should be combined.

Given S and A, our system looks for possible entailment relations between various fragments of S and A, namely words, bigrams, trigrams, and parse tree fragments. To assess entailment between these components, the system considers equality, synonymy, and entailment relationships drawn from three sources: WordNet (hypernyms), the DIRT paraphrase database, and the ParaPara paraphrase database from Johns Hopkins University (described shortly). Each relationship found is a piece of evidence that S entails A. Then, the various evidence is combined to provide an overall confidence in entailment. For this task, the absolute confidence number is not important, as our algorithm only cares about the *relative* entailment strength in order to find the sentences that most likely entail A.

Word and N-gram Matching:

The simplest entailment is word matching: a word in S matches (i.e., is identical to, taking into account morphological variants) a word in A, e.g., “produce”(in S) →

“production”(in A); "suppporting"(S) → "supportive"(A). Word matches are scored according to the word’s "importance", i.e, the extent to which it carries the semantic meaning of the phrase or sentence in which it occurs. For example, words like "HIV", "virus", "infection" (for the AIDS topic) carry more of the meaning than general words like "go", "move", etc. To capture this intuition we use two measures of "importance":

- **Salience:** an Idf (inverse document frequency) measure of how unlikely a word is, with uncommon words being weighted higher than common words, defined as:

  salience(w) = max [log (1/p(w|topical-documents)) - k, 0]

  k is a constant chosen such that a pool of very common words ("of", "the", "a") have a salience of 0. A word has high salience if it occurs infrequently, and the most common words have a salience of 0.

- **Topicality:** A word that occurs unusually frequently for documents about a particular topic (relative to general text) is considered topical, and can be given more weight. We define it as:

  topicality(w) =
  max[log (p(w|topical-documents)/p(w|general-documents)) - 1, 0]

  A word has high topicality if it occurs unusually frequently in domain texts (relative to general texts), while a word that is no more/less frequent in domain texts than general texts (independent of its *absolute* frequency) has a topicality of 0. Topicality helps to distinguish between domain-general and domain-specific terms, allowing us to place more weight on domain-specific terms relative to equally infrequent domain-general terms (e.g., weight "virus" more than "primarily" for documents about AIDS).

The overall entailment strength is a weighted combination of these measures:

weight(w) = λ.topicality(w) + (1- λ).salience(w)

where λ controls for the relative weights of topicality and salience. p(w|topical-documents) is estimated from the QA4MRE background collection for the topic of the question (AIDS, climate change, music and society, Alzheimer), and p(w|general-documents) is estimated from the British National Corpus (BNC Consortium, 2001). For Ngrams, we add the weights of individual words in the Ngrams. We optimized for λ on the 2011 QAMRE data, producing an optimal value of λ = 0.9.

Use of Paraphrases to Identify Entailments:

In addition to looking for exact word/phrase matches, the system also looks for entailment relations using two paraphrase databases, namely ParaPara (Chan et al., 2011) and DIRT (Lin and Pantel, 2001):

- The ParaPara paraphrases are of the form $string_1 \rightarrow string_2$ (string equivalents), found using a combination of bilingual pivoting and monolingual distributional similarity. Bilingual pivoting uses aligned parallel corpora in multiple languages. If a phrase X is aligned with a foreign phrase Y, and that foreign phrase Y is aligned with a phrase Z, then X and Z may be paraphrases. The set of high confidence paraphrases found this way is then given a second score based on the distributional similarity of their surrounding words in the Google N-gram corpus. Finally, the two scores are combined together using an SVM trained on human-annotated training data. Some of the better examples that applied in QA4MRE are:

  "total lack of" → "lack of"
  "increasingly" → "every day"
  "cause" → "make"
  "cooperation with" → "closely with"
  "finance" → "fund the"
  "against aid" → "fight aid"
  "to combat" → "to fight"
  "spend" → "devote"
  "one" → "member"
  "reason" → "factor"
  "is one of" → "among"

- The DIRT paraphrases are of the form **IF** (X r_1 Y) **THEN** (X r_2 Y), where r_1 and r_2 are dependency paths between words X and Y, and are based on distributional similarity: (X r_1 Y) and (X r_2 Y) are paraphrases if the frequency distribution of the Xs with r_1, and of the Xs with r_2, are similar (combined with the same for the Ys). For our purposes here we simply use the **IF** (X ←*subj*– $verb_1$ –*obj*→ Y) **THEN** (X ←*subj*– $verb_2$ –*obj*→ Y) paraphrases as string equivalents ($verb_1 \rightarrow verb_2$), although with more engineering we could also use longer paraphrases in the database and exploit the dependency structure more. In our earlier work, we found these simple verbal paraphrases to be the most reliable. Some of the better examples that applied in QA4MRE are:

  IF X decreases Y **THEN** X reduces Y
  IF X increases Y **THEN** X grows Y
  IF X offers Y **THEN** X includes Y
  IF X names Y **THEN** X calls Y
  IF X supports Y **THEN** X funds Y
  IF X causes Y **THEN** X affects Y

If a paraphrase X → Y (from either ParaPara or DIRT) applies during the entailment assessment, then we score the entailment as weight(x)*confidence(paraphrase), where the confidence(paraphrase) is provided in the paraphrase databases (range 0-1). In other words, we downgrade the strength of an exact match (X → X) by confidence(paraphrase). If the paraphrase occurs in both databases we take the highest

confidence for it. This formula was found to be the most effective in the trials with ran.

WordNet

We also use WordNe in a straightforward way to find word hypernyms and synonyms. As well as using the WordNet synsets, we also use WordNet's pertainym ("pertains to") links. Pertainyms cross part-of-speech boundaries and offer some useful additional synonyms to the synsets, for example the pertainyms of (senses) of "quick" are:

> "quickly" "speedily" "quicker" "faster" "quickly" "quickest"
> "fastest" "prompt" "quick" "quickly" "agilely" "nimbly"

Syntactic Fragments

In addition to word and phrase-based matching, we also match parse tree fragments from parses of the two phrases/sentences. This allows some credit to be given if syntactic dependencies between words is preserved between the two sentences. We use (a modern version of) the SAPIR parser (Harrison and Maxwell, 1986), a broad coverage chart parser, generate a dependency-like structure from the parse, and then "shred" the dependency tree into fragments, where each fragment denotes one dependency-style link. If a sentence S and answer A share a syntactic fragment, then this constitutes another piece of evidence that S entails A.

2.3 Task 2: Implication Assessment (Proximity Detection)

Given the system has identified sentences in the document that most likely express Q and A in the document, the second task is to assess how likely it is that the combination of these sentences imply that A is an answer to Q. Ideally we would find some appropriate syntactic connection between those two sentences (e.g., "Q-sentence *because* A-sentence"). However, this task is difficult because the connection may be indirect or simply not stated explicitly in the text. To deal with this, we perform task 2 in a crude way, by measuring the distance between the sentences, prefering the closest pair. In the case of a tie, we then prefer the pair that also most strongly entails Q and A, using the scores from Task 1.

Algorithmically, we:

1. Find the N sentences SQ_i that most strongly entail Q
3. Find the N sentences SA_{nj} that most strongly entail A_n, for each of the five candidate answers A_n
4. Find the pair $<SQ_i,SA_{nj}>$ where the distance (number of sentences) between SQ_i and SA_{nj} in the document is smallest. For ties, prefer the pair where the entailments of $SQ_i \rightarrow Q$, and $QA_{nj} \rightarrow A$ are strongest.

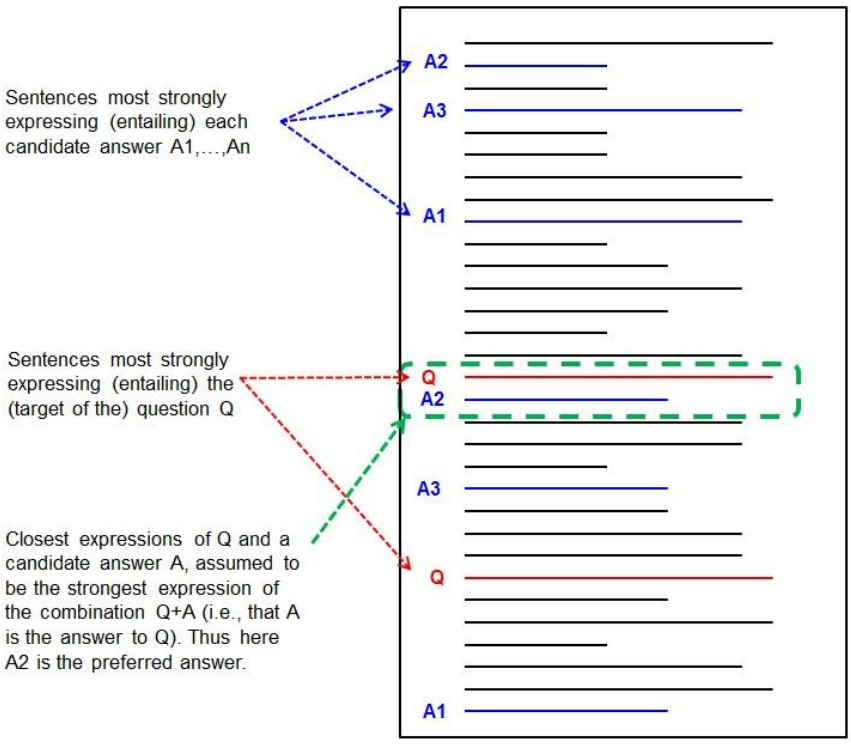

Fig. 1. The system finds the closest pair of sentences, one plausibly entailing Q, one plausibly entailing an answer A_i, and concludes the answer is A_i

Experimentally, we used N = 3, as it achieved the highest performance on the 2011 training data. Figure 1 illustrates this process.

2.4 Example

As an example of the system's behavior consider the following question:

Q[3.5] What is one of the MCP goals in Third World countries?
A1: funding international organizations
A2: separation of HIV-positive mothers from their infants
A3: elimination of the contribution to the Global Fund
A4: participation in USAID
A5: separation of family planning from HIV prevention [CORRECT]

First the system finds the top three sentences most likely entailing Q and each answer A_i, as illustrated in Table 1:

Question Q / Candidate answer Ai	3 sentences that most strongly entail Q or Ai
Q[3.5] What is one of the MCP goals in Third World countries?	S7 S30 S52
A1: funding international organizations	S2 S29 S35
A2: separation of HIV-positive mothers from their infants	S2 S6 S10
A3: elimination of the contribution to the Global Fund	S26 S29 S30
A4: participation in USAID	S13 S29 S34
A5: separation of family planning from HIV prevention [CORRECT]	S30 S31 S45

Table 1. First the system finds the 3 sentences that most likely entail Q, and for each A_i the 3 sentences that most likely entail that A_i

For example, there is strong evidence that S30 likely entails Q due to the paraphrases "aimed at" → "goal" and "developing countries" → "Third World countries" (both from ParaPara):

Q[3.5] What is one of the MCP **goals in Third World countries**?
S30: ...U.S. funding...will be saddled by...restrictions **aimed at** separating family planning from HIV prevention **in developing countries**.

Let SQ be the set of 3 sentences most entailing Q, i.e., {S7,S30,S52}, and SA be the set of 5x3=15 sentences most entailing one of the A_i, i.e., {S2,S6,…,S35,S45}. Next the system looks for $<S_Q \in SQ, S_{Ai} \in SA>$ pairs from these sets with the mimimum (sentence number) distance between some entailing sentences (3x3x5 pairs to consider). There are two pairs where the distance is zero (the same sentence entails both Q and an A_i):

Q+A3, both plausibly entailed by S30 (see Table 1)
Q+A5, both plausibly entailed by S30 (see Table 1)

To break the tie, the system looks at the strengths of the entailments. Using the scoring metric earlier, the scores are:

For A3: 7.36 (S30→Q) + 10.85 (S30→A3) = 18.21
For A5: 7.36 (S30→Q) + 33.48 (S30→A5) = 40.84

Thus answer A5 ("separation of family planning from HIV prevention") is selected (in this case this is the right answer). The reason the entailment strength is so high (33.48) for this entailment is obvious, as S30 contains A5 almost verbatim within it:

A5: **separation of family planning from HIV prevention**
S30: ...U.S. funding...will be saddled by...restrictions aimed at **separating family planning from HIV prevention** in developing countries.

Note that the entailment S30 → Q+A5 is still only partial; for example the system did not find evidence related to MCP in the question, i.e., it did not prove that the goals in S30 were "MCP goals" (Q). More generally, there are typically elements of S and A that are left unentailed. However, for the QA4MRE task, we only need to know the relative entailment strengths in order to select the best answer.

3 Experimental Results

We submitted three runs with different parameters for scoring, the highest run achieving an accuracy of 0.40 (versus a baseline of random guessing of 0.20). The c@1 score is also 0.40, as the system always guesses an answer.

We also performed some ablation studies to see the effects of adding/removing different knowledge sources. The results are shown in Table 2, using the current version of system:

Subtractive ablations	
42.5	Main system (all resources)
41.9	minus WordNet (only)
38.1	minus ParaPara (only)
41.9	minus DIRT (only)
38.1	baseline (none of the resources)
Additive ablations	
38.1	baseline (none of the resources)
41.9	add WordNet (only)
39.4	add ParaPara (only)
41.9	add DIRT (only)
42.5	Main system (all resources)

Table 2. Precision (percent correct). The knowledge resources contribute approximately 4% to the system's accuracy (also c@1) on the 2012 QA4MRE data.

The patterns in the ablation results are somewhat varied, illustrating the interactions that can occur between the scores from different knowledge resources, and making it difficult to draw detailed conclusions about individual resources. However, the general picture from these studies is clear: the basic (resource-free) algorithm accounts for the majority of the score (38%), with the knowledge resources together contributing an additional 4% to the score, and with no single knowledge resource clearly dominating the other.

4 Discussion

There are clearly many improvements that can be made. We summarize some of these here, before finally turning to the larger challenge of Machine Reading.

4.1 Better question analysis

We did not make any attempt to analyze the questions beyond extracting words, bigrams, and parse fragments, although clearly knowing the question type would affect what kind of an answer is sought. In addition, in one case, there is no domain-specific content to the questions at all:

Q[12.12]: Which of the following is true?

Here it is pointless to search for sentences entailing (the target of) question Q[12,12] (as there is no target), and the results will be random. Better analysis of the questions, in particular identification of the question type, would help improve performance.

4.2 Sentence-level entailments and anaphora resolution

Although we allow the content of the Q+A pair to be split over multiple sentences, we assume that the semantic content of Q alone, and A alone, is within a single sentence. In practice, this assumption does not always hold, for example a pronoun may refer back to previous sentences (our system does not currently do anaphora resolution). For instance, in this example:

Q[3.9] Why did Burney decide not to write a biography of Dr Samuel Johnson?
S44: At one time he thought of writing a life of his friend Dr Samuel Johnson, but he retired before the crowd of biographers who rushed into that field.

our system's inability to realize that "he" (in S44) is "Burney" (mentioned in earlier sentences) weakens the assessed entailment between S44 and Q. Our system gets this question wrong.

In addition, several questions and answers themselves use pronouns, e.g.,:

Q[9.9] How did he earn his degrees at Oxford?
1. He wrote an essay on comets
2. He produced an operetta at Drury Lane
3. He studied with Dr Arne
4. He composed various pieces
5. He sang in a choir

Without identifying who "he" is (which our system does not do), our entailment reasoning is missing a critical piece of information to be entailed, again weakening its entailment reasoning.

4.3 Proximity as an Approximation for Capturing Q+A Together

Our method assumes that if two sentences S_Q and S_A expressing Q and A respectively are close, then the two sentences together likely expresses the combined meaning of Q+A. Although this is clearly a gross assumption, it appears to holds true surprisingly often. The larger problem we observe is that, as currently implemented, proximity always takes precidence over the strengths of entailments. A bad example is as follows:

Q[2.8] What advantage does the Jatropha curcas offer?
A2: it grows on semi-arid fields [actual correct]
A4: it reduces pests of other crops [predicted correct]

The system selects A4 because a sentence S16 entails both Q and A4 with moderate strength. However, inspection of the data shows that two other sentences, S2 and S3, very strongly entail A and Q respectively:

S2: It sounds too good to be true: a biofuel crop that **grows on semi-arid lands** and degraded soils...
S3: That is what some are claiming for **Jatropha curcas**, the miracle biofuel crop.

Despite the relatively strong entailments, our system disprefers this option as the two sentences are further apart (distance 1, rather than distance 0), and sentence proximity currently takes absolute priority over entailment strengths once the top 3 entailing sentences for each answer have been selected. In future, we could consider a weighted combination of the distance and entailment strength metrics when selecting an answer.

4.4 Short answers

When an answer is short, e.g., a single word, there is very little to be "entailed" by the text. In the worst case, if the answer word is common then there is little basis to pick the best 3 sentences that "entail" that answer. Our system did notably worse on questions with single-word answers, for example:

Q[2.5] Which is the biofuel whose production most reduces the emission of greenhouse effect gases?
1. bio-diesel
2. bio-oil
3. corn ethanol
4. cellulosic biofuel
5. gasoline

In the supporting document, 9 sentences contain the phrase "corn ethanol", and thus there is no basis (using our current algorithm) to select the 3 sentences that "most

strongly entail" corn ethanol from them. Again, a modification of the algorithm to allow more entailing sentences in its set of candidates could overcome this problem.

5 Towards Machine Reading

Finally we consider the larger goal of Machine Reading, and the architecture of our system within it. Although our system's performance was relatively respectable, there is still a long way to go. Most significantly, our system is largely relying on local entailment estimations, and does not make any attempt to construct an overall model of the text, resulting in sometimes brittle behavior, for example when there are many word-level entailments between the text T and the hypothesis H, but the relationships between those words in T and between those words in H are completely different.
There are two major challenges to overcome to move closer towards machine reading: the knowledge problem, and the reasoning problem. The **knowledge problem** is that large amounts of world knowledge are needed to fully identify entailments, but our existing resources (e.g., paraphrase databases) are still limited. Although paraphrasing allows some simple entailments to be recognized, e.g. **IF** X contains Y **THEN** Y is inside X, there are many cases in the QA4MRE dataset where the gap between T and H is substantial. Some examples of semantically similar, but lexically different, phrasings are shown below requiring considerable knowledge to recognize the approximate equivalence:

Q[2.7] What is the **external debt** of all African countries?
S61 Africa **owes foreign banks and governments** about 350 billion.

Q[2.1] When did the rate of AIDS started to **halve** in Uganda?
S73 The rate of AIDS in Uganda is **down to about 8, from a high of 16** in the early 1990s.

A4[to Q7.9] to **encourage the use of** groundwater
S73 ...the UN has ...a program to **give them access to** groundwater sources.

Clearly more work is needed to acquire lexical and world knowledge to help recognize such near-equivalences.
Concerning reasoning, text allows many weak entailments to be made, but the **reasoning problem** is how to combine all those entailments together into a coherent whole. This is something our system does not do; given some text, it can posit many weak entailments, many of which are contradictory, but does nothing to try and find a subset of those entailment which are together coherent. One can view this as the challenge of "reasoning with messy knowledge". Part of the challenge is devising a suitable method for reasoning with uncertainty, so that contradictions in the entailments can be best resolved. (A promising candidate for this is Markov Logic Networks (Richardson and Domingos, 2006)). However, simply ruling out inconsistencies may

not be a sufficient constraint on the problem; When people read, they also bring large-scale expectations about "the way the world might be", and coerce the fragmented evidence from the text to fit those expectations. Reproducing this kind of behavior computationally has long been a goal of AI (Minsky, 1974; Schank and Abelson, 1977), but still remains elusive, both in acquiring such expectations and using them to guide reading (Clark, 2010). Recent work on learning event narratives ("scripts"), e.g., (Chambers and Jurafsky, 2008) and building proposition stores (Van Durme et al., 2009; Penas and Hovy, 2010) offers some exciting new possibilities in this direction.

6 Summary

Our system for QA4MRE is based on assessing entailment likelihood, and breaks down the problem into two parts:

i. finding sentences that most entail the question Q and each answer A_i
ii. finding the closest pair of such sentences where one entails Q and the other A_i.

The system's best run scored 40% correct. As just discussed, to progress further, the system needs to move from assessing local entailments to constructing some kind of "best coherent model", built from a subset of those (many) weak entailments. This requires both addressing the knowledge problem (of acquiring the knowedge to support that) and the reasoning problem (to construct such a model). The QA4MRE challenge itself, though difficult, is one that seems ideally suited to promoting research in these directions.

References

BNC Consortium. *The British National Corpus*, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Chambers, N., Jurafsky, D. "Unsupervised Learning of Narrative Event Chains", in Proc ACL'08, 2008.

Chan, T., Callison-Burch, C., Van Durme, B. *Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity*. EMNLP Workshop: GEMS. 2011

Clark, P. Do Scripts Solve NLP? Working Note 28, AI Lab, Univ Texas at Austin. 2008.

Graesser, A. C. 1981. "Prose Comprehension Beyond the Word". NY:Springer.

Harrison, P., Maxwell, M. "A New Implementation of GPSG", Proc. 6th Canadian Conf on AI (CSCSI'86), pp78-83, 1986.

Lin, D., and Pantel, P. "Discovery of Inference Rules for Question Answering". Natural Lan-guage Engineering 7 (4) pp 343-360, 2001.

MacCartney, B. Natural language inference. Ph.D. dissertation, Stanford University, June 2009

MacCartney, B., Manning, C. Natural logic for textual inference. ACL Workshop on Textual Entailment and Paraphrasing, Prague, June 2007

Minsky, M. A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306, June, 1974.

NIST. Proceedings of the fourth Text Analysis Conference (TAC'2011), 2011.

Richardson, M., Domingos, P. Markov Logic Networks. Machine Learning 62:107-136. 2006.

Peñas A., Hovy, E. Filling Knowledge Gaps in Text for Machine Reading. 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, 2010.

Schank, R., Abelson, R. Scripts, Plans, Goals and Understanding. NJ: Erlbaum. 1977.

Van Durme, B., Michalak, P., Schubert, L. "Deriving generalized knowledge from corpora using WordNet abstraction", 12th Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL-09), 2009.