Answering and Questioning for Machine Reading

Lucy Vanderwende

Microsoft Research
Redmond, WA 98052
lucyv@microsoft.com

Abstract

Machine reading can be defined as the automatic understanding of text. One way in which human understanding of text has been gauged is to measure the ability to answer questions pertaining to the text. In this paper, we present a brief study designed to explore how a natural language processing component for the recognition of textual entailment bears on the problem of answering questions in a basic, elementary school reader. An alternative way of testing human understanding is to assess one's ability to ask sensible questions for a given text. We survey current computational systems that are capable of generating questions automatically, and suggest that understanding must comprise not only a grasp of semantic equivalence, but also an assessment of the importance of information conveyed by the test. We suggest that this observation should contribute in the design of an overall evaluation methodology for machine reading.

Introduction

Machine reading can be defined as the automatic understanding of text. The question being raised in this workshop is how to leverage the tools that have been developed in Natural Language Processing (NLP), i.e., parsing, semantic role labeling and text categorization, and the recent advances in machine learning and probabilistic reasoning, in order to improve automatic text understanding. The question not raised in the workshop proposal, however, is how to evaluate whether and when automatic text understanding has in fact improved.

One obvious method of evaluating machine reading is to test the machine as we do humans, by determining how well a system performs on a standard set of test questions. Such an approach was explored in Project Halo (Friedland et al. 2004). Focusing on the NLP tools developed to handle textual entailment, we will describe a brief study intended to test how useful such a tool is in accomplishing what should be a simple task for a reading machine: answering questions from a grade school reading book. Answering questions at this grade level is interesting because the questions do not presuppose extensive world knowledge, but rather, the questions often test common-sense knowledge, keeping the task for machine reading focused on reasoning capabilities rather than on the acquisition of complex bits of world knowledge.

We will first describe the tools that we have developed for handling textual entailment, a component which is typically assumed to be an integral part of text understanding; these tools leverage both NLP and machine learning.

While the tools we have built are among systems that perform very well on a standard test suite for textual entailment, we note that the ability to answer first grade reading questions is passable (given correct anaphora resolution), but the ability to answer questions at grade level 2 is already poor. This is due in part to the fact that the material for answering the questions at level 2 typically span several sentences, and in part due to the need for extensive anaphora resolution, well beyond pronominal coreference, and the recovery of understood arguments.

We will then discuss two additional areas which are of particular interest to us in the area of machine reading: machine-generated questions and the evaluation of machine reading. These two areas can be viewed as related in the following way: text generation algorithms allow many questions to be generated for any given input sentence, but many, if not most, of these do not strike a reader as reasonable questions. We therefore propose that one avenue for evaluating machine reading might include an evaluation of the questions that are automatically generated. Questions generated should be expected to be reasonable, thus demonstrating some level of human understanding, and not merely a mechanical repetition of the input. Such an evaluation method would reward systems that attempt to synthesize information from multiple sentences and/or sources and that develop a set of beliefs representing a target audience knowledge.

Basic Semantic Understanding: Textual Entailment

Recognizing the semantic equivalence of two fragments of text has proven to be both a critical component of processing natural language text and a great challenge for NLP. In the recent PASCAL Challenge Recognizing Textual En-

tailment (RTE)[1], this task has been formulated as the problem of determining whether some text sentence T entails some hypothesis sentence H (see Dagan et al. 2005). Some examples of Text and Hypothesis sentences are:

> T: Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.
> H: Yahoo bought Overture.
> True

> T: Microsoft's rival Sun Microsystems Inc. bought Start Office last month and plans to …
> H: Microsoft bought Star Office.
> False

> T: Since its formation in 1948, Israel fought many wars with neighboring Arab countries.
> H: Israel was established in 1948.
> True

> T: The National Institute for Psychobiology in Israel was established in May 1971 as …
> H: Israel was established in 1971.
> False

As these examples show, correctly identifying whether two text segments are semantically equivalent or not is one necessary facet of human or human-like understanding. Another fundamental aspect of understanding, the ability to distinguish important vs. background, will be explored in a later section.

MENT: Microsoft Entailment

Given that the RTE task presents us with pre-selected pairs of sentences, MENT takes the approach of predicting false entailment rather than attempting to predict true entailment. This approach is motivated by our earlier observations (Vanderwende and Dolan 2005) that twice as many RTE test items could be determined to be False, than True, using syntax and thesaurus. MENT begins with logical form representation of both text and hypothesis sentences, in which all the relations between syntactic dependencies have been labeled. Our algorithm proceeds as follows (described in detail in Snow et al. 2006):

1. Parse each sentence, resulting in syntactic dependency graphs for the text and hypothesis sentences.
2. Attempt an alignment of each content node in the dependency graph of the hypothesis sentence to some node in the graph of the text sentence
3. Using the alignment, apply a set of syntactic heuristics for recognizing false entailment; if any match, predict that the entailment is false.
4. If no syntactic heuristic matches, back off to a lexical similarity model, with an attempt to align detected paraphrases.

This system was among the top performing systems in RTE-2 (see Bar-Haim et al. (2006), for details) and so is representative of the state-of-the-art in systems attempting to recognize textual entailment.

Reading Comprehension

In this section, we describe a brief study designed to explore the extent to which a textual entailment component is useful in the task of answering the questions in a grade school reader. A priori, we do not expect to need access to extensive world knowledge, though the questions show that commonsense knowledge and reasoning is very much targeted in these exercises.

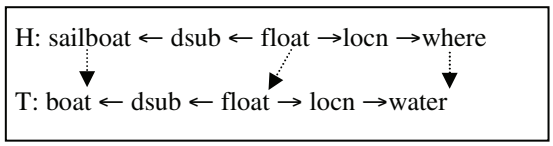
Consider a story and questions targeting reading comprehension at grade 1 level (Reading Comprehension, grade 1, 1998, page 77):

> See the boats! They float on water. Some boats have sails. The wind moves the sails. It makes the boats go. Many people name their sailboats. They paint the name on the side of the boat.
>
> Questions:
> 1. What makes sailboats move?
> 2. Where do sailboats float?

As in RTE, we will manually choose the sentence pairs to submit to the textual entailment system. The system predicts the correct answer for 2.

T: They float on water.
H: Where do sailboats float?

MENT computes the logical forms for each of these two sentences (including basic anaphora resolution so that "they" is coreferent with "boats") and computes the alignment of the nodes in the hypothesis sentence to the nodes in the text sentence, as show below:



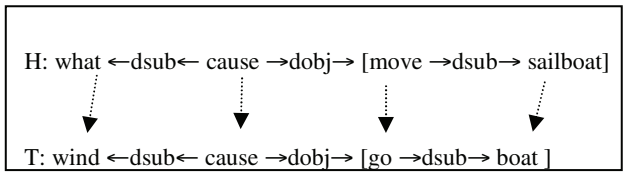
[1] http://www.pascal-network.org/challenges/RTE.

No rules for detecting false entailment match, and so MENT predicts True, indicating that this sentence contains the correct answer. Note that in addition to textual entailment, simple pronominal coreference and lexical synonymy were required in order to find the correct alignments.

For question 1, we chose the following pair of sentences to submit to the MENT system:

T: It makes the boats go.
H: What makes sailboats move?

For this pair of sentences, the correct alignment can still be established, and so the answer can be predicted correctly, but more nodes require lexical synonym lookup, as shown below:



Now consider a different story, for grade 2, which is specifically targeting the young reader's ability to make inferences (Reading Comprehension, grade 2, 1998):

> A father sea horse helps the mother. He has a small sack, or pouch, on the front of his body. The mother sea horse lays the eggs. She does not keep them. She gives the eggs to the father.
>
> Questions:
> 1. What does the mother sea horse do with her eggs?
> 2. Where does the father sea horse put the eggs?

We choose the following optimal pairs of sentences for questions 1 and 2. For question 1, MENT is able to predict that this sentence is a good answer to the question.

Question 1
T: She gives the eggs to the father.
H: what does the mother sea horse do with her eggs?

Question 2:
T: He has a small sack, or pouch, on the front of his body.
H: Where does the father sea horse put the eggs?

For question 2, however, MENT rejects the pair as False since there is no match for the verb associated with the aligned subject of the hypothesis (see Snow et al. (2006) for details concerning heuristics), "having" and "putting" are not synonyms. Furthermore, there is no match for the object in the Hypothesis, "eggs". This type of question, which involves a chain of reasoning, is not appropriately handled by textual entailment. Commonsense information of the type that can be extracted from dictionaries (see Vanderwende 1995) plays a role in detecting the answer[2], but basic reasoning capabilities are needed to further align these sentences and discover the answer.

Note that it would be possible to search for the answer on the WWW and discover the correct answer by mining the information across a sizable set of sites, an approach which would likely be taken by a typical Question-Answering system. However, given the expectation that a second grader can correctly answer question 2 using only this text, without recourse to external materials, we would like to suggest that an information-mining approach to machine reading does not actually address the core problem of machine reading, as the following example illustrates.

Consider this example from a second grade reader[3]:

> One night, not very far from here, a fox was looking for food … But the sad thing was that there was no food to be seen.
> The fox was getting very hungry.
> "If I don't find something to eat soon," said the fox to himself, "I shall die of hunger".
> The fox looked in the bushes. He looked behind the trees. He looked in the fields. He looked everywhere. Then he came to a well.
>
> Question: Where did the fox look for food?

As proficient readers, we can easily answer this question: in the bushes, behind the trees, in the fields, and in the well. We choose the following pair of sentences to submit to the entailment system.

T: The fox looked in the bushes.
H: Where did the fox look for food?

MENT predicts False, since there is a concept in the hypothesis that is not aligned to any concept in the text, namely "food". This can be overcome by a system is capable of recovering understood arguments.

---

[2] The information that "sack" and "pouch" are the typical subject of activities such as "hold, keep, carry" can be automatically identified from various dictionary definitions.
[3] Excerpted from: The Fox and the Wold, Horsburgh, 2003.

The story continues:

> The fox walked round the well and looked in. Suddenly, he stopped. There, in the water, was the reflection of the moon. "Ho, ho, ho!" he laughed. "What a luckily little fox I am! Here I am thinking about buckets and ropes when in front of my eyes is some food. Someone has thrown a whole cheese into the well …"
> The silly fox did not know that the cheese was not cheese. It was the reflection of the moon.

The student must now answer these questions:

> What did the fox think there was in the well?
> What did the fox really see in the well?

Even if we manually select the pairs of sentences to submit to the textual entailment, it is not clear what the optimal set of pairs is. Below are some possibilities, though no pairing will product the correct answer.

T: … in front of [fox's] eyes is some food.
H: What did the fox think there was in the well?

T: Someone has thrown a whole cheese into the well.
H: What did the fox think there was in the well?

T: It was the reflection of the moon.
H: What did the fox really see in the well?

What the appropriate steps in the reasoning process might be exactly is a matter for investigation, but it is certainly clear that in this case, we must rely solely on the interpretation of this text in order to answer the question correctly; no amount of searching or pre-processing of the web will produce the correct answer.

The importance of being important: Question generation

Another method of testing reading comprehension and of improving human learning is by means of question asking. This method of testing confirms that students have identified the main ideas and indicate which part of the learning material is important and worth testing (see, e.g., Chang et al. 2005). In this section, we describe several computer systems which have the capability of generating questions automatically. A review of these systems makes clear that the process of asking good questions, and identifying which part of the text is worth generating questions about, is still a real challenge.

Ureel et al., 2005, build on the tradition of the Socratic method of learning by creating a computer system, *Ruminator*, that reflects on the information it has acquired and poses questions in order to derive new information. Their hypothesis is that "a system that can generate its own questions to work on will be able to learn the ways in which new information fits with existing knowledge, detect missing knowledge, and provide better explanations than systems without this ability" (Ureel et al. 2005). *Ruminator* takes as input simplified sentences in order to focus on question generation rather than handling syntactic complexity, and is capable of asking typical Journalists' questions, who, what, when, where, why and how. It is unclear whether *Ruminator* can ask questions that span more than one sentence, but it is reported that even a single sentence generated 2052 questions. The authors note that it is important "to weed out the easy questions as quickly as possible, and use this process to learn more refined question-posing strategies to avoid producing silly or obvious questions in the first place" (Ureel et al. 2005). Furthermore, some of the examples of automatically generated questions provided are semantically ill-formed and answering these would presumably not lead to any enhanced knowledge: "Is it true that the City of San Antonio's spouse is Chile?" (Ureel et al. 2005). A key component that appears to be missing from the system design is an estimation of the utility, or informativeness, of an automatically generated question.

Mitkov and An Ha (2003) describe a computer system for generating multiple-choice questions automatically, "based on the premise that questions should focus on key concepts rather than addressing less central and even irrelevant concepts or ideas" (Mitkov and An Ha 2003). In order to accomplish this, the system comprises a set of transformational rules, a shallow parser, automatic term extraction and word sense disambiguation. Questions are only asked in reference to domain-specific terms, to ensure that the questions are relevant, and sentences must have either a subject-verb-object structure or a simple subject-verb structure, which limits questions to core concepts in the sentence; naturally, questions are only generated one sentence at a time. They tested this method on a linguistics textbook and found that 57% were judged worthy of keeping as test items, of which 94% required some level of post-editing. 43% of the automatically generated questions were rejected as either not focusing on a central concept, or requiring too much post-editing. One example of a generated question is (Mitkov and An Ha 2003):

> "which kind of language unit seem to be the most obvious component of language, and any theory that fails to account for the contribution of words to the functioning of language is unworthy of our attention"

This question was considered worthy, after revising the question by eliminating the second coordinate constituent.

Schwartz et al. (2004) describe a system for generating questions which also comprises the NLP components of lexical processing, syntactic processing, logical form, and

generation. The input to the generation component is a logical form, and from this logical form the typical Journalist questions can be generated. This system uses summarization as a pre-processing step as a proxy for identifying information that is worth asking a question about. Questions can be generated for any constituent, including prepositional phrases. For the input sentence "At school, John eats rice every day", a number of questions can be generated, among which "Where does John eat rice every day?" Nevertheless, the authors note that "limiting/selecting questions created by Content QA Generator is difficult" (Schwartz et al. 2004).

Finally, Harabagiu et al. (2005) describes an approach to automatic question generation using syntactic patterns, together with semantic dependencies and named entity recognition. Their purpose was "to generate factoid questions automatically from large text corpora" (Harabagiu et al. 2005). User questions were then matched against these pre-processed factoid questions in order to identify relevant answer passages in a Question-Answering system. While no examples of automatically generated questions are provided, this study does report a comparison of the retrieval performance using only automatically generated questions and manually-generated questions: 15.7% of the system responses were relevant given automatically generated questions, while 84% of the system responses were deemed relevant with manually-generated questions. The discrepancy in performance indicates that significant difficulties remain.

Questioning as Evaluation

The studies discussed above show that identifying which text segments are important and worth asking questions about remains a great challenge for computational systems. From the few examples presented in these studies, it appears easy for humans to judge whether an automatically generated question is sensible and not merely mechanical, obvious and/or nonsensical.

Given that machine reading, like human reading, should include an understanding of what information is important and in focus for a particular text, and given that adequate question generation is taken as a demonstration of understanding and will lead to enhanced learning, then we might consider how to incorporate question generation in an evaluation methodology for machine reading.

Evaluating questions will necessarily involve human judgment, given the many questions that can sensibly be asked for a given text, though human judgment is very quick and should achieve high agreement on this task. A human-in-the-loop has the additional disadvantage that the evaluation metric cannot be used during system development, which is generally considered necessary for training machine-learned algorithms. These disadvantages should be weighed against nourishing research directed at establishing the importance of information conveyed by text rather than treating the text as a great morass of facts all of which are equally important.

Conclusion

We hope to have shown that while textual entailment is a basic component for understanding the semantic content of a given text, it alone is not sufficient for the task of answering questions in an elementary grade reader. In order to accomplish this, systems must expand their capabilities to advanced anaphora resolution, recovering understood arguments, and components that begin to identify the necessary reasoning steps. The advantage of focusing on answering questions in basic readers is that very little external world knowledge is assumed, and the focus is rather on commonsense knowledge.

We discussed that a complementary and/or alternative way to test human-level understanding of a given text is to evaluate how sensible automatically-generated questions are. Current systems have demonstrated the ability to generate questions, but all systems suffer in over-generating questions, i.e., generating questions that a human would easily reject as being nonsensical. For this reason, we suggest that at least one component of an evaluation methodology for machine reading should be an evaluation of the quality of questions. This will lead us to research leveled at determining the important information that humans derive from a given text, and not only a rote memorization of the facts that the text represents. Question generation tests the identification of important information, one aspect of human reading and understanding, in a way that question answering does not.

References

Bar-Haim, R.; Dagan, I.; Dolan, W.B.; Ferro, L.; Giampiccolo, D.; Magnini, B.; and Szpektor, I. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment.

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science, Volume 3944, Jan 2006*, Pages 177-190.

Friedland, N.; Allen, P.G.; Witbrock, M.; Matthews, G.; Salay, N.; Miraglia, P.; Angele, J.; Staab, S.; Israel, D.; Chaudhri, V.; Porter, B.; Barker, K.; and Clark, P. 2004. Towards a Quantitative, Platform-Independent Analysis of Knowledge Systems, In *The Ninth International Confer-*

ence on the Principles of Knowledge Representation and Reasoning (KR2004)

Harabagiu, S.; Hickl, A.; Lehmann, J.; and Moldovan, D. 2005. Experiments with Interactive Question-Answering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL05).

Horsburgh, N. 2003. *Oxford Reading Circle, book 2*. Oxford University Press, New Delhi, India.

Mitkov, R., and An Ha, L. 2003. Computer-Aided Generation of Multiple-Choice Tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, Edmonton, Canada, May, pp. 17 – 22

Reading Comprehension, grade 1. 1998. School Specialty Publishing.

Reading Comprehension, grade 2. 1998. School Specialty Publishing.

Schwartz, L.; Aikawa, T.; and Pahud, M. 2004. Dynamic Language Learning Tools. In *Proceedings of the 2004 InSTIL/ICALL Symposium*, June 2004.

Snow, R.; Vanderwende, L.; and Menezes, A. 2006. Effectively using syntax for recognizing false entailment. In Proceedings of HLT/NAACL 2006.

Ureel II, L. C.; Forbus, K.; Riesbeck, C.; and Birnbaum, L. 2005. Question Generation for Learning by Reading. In the Proceedings of the AAAI Workshop on Textual Question Answering, Pittsburgh, Pennsylvania. July 2005

Vanderwende, L., and Dolan, W.B. 2006. What syntax can contribute in entailment task. In *Lecture Notes in Computer Science, Volume 3944, Jan 2006*, pp. 205-216.

Vanderwende, L. 1995. The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries. Ph.D. thesis, Dept. of Linguistics, Georgetown University, Washington, D.C.