Question Generation: Papers from the 2011 AAAI Fall Symposium (FS-11-04)

Generating More Specific Questions

Xuchen Yao
Department of Computer Science
Johns Hopkins University

Abstract

Question ambiguity is one major factor that affects question quality. Less ambiguous questions can be produced by using more specific question words. We attack the problem of how to ask more specific questions by supplementing question words with the hypernyms for answer phrases. This dramatically increases the coverage of generated which questions. Evaluation results show improved question quality when the question words are disambiguated correctly given the context.

1 Introduction

Question Generation (QG) is the task of generating reasonable questions from an input, such as a text or a database. The generated question is either displayed to human beings for mostly educational purposes (such as testing students' understanding of a text) or sent to computer input for further processing (such as building an FAQ index). High quality questions usually target on asking the key piece of information from the input while still maintaining grammaticality and naturalness.

In practice, however, computers sometimes generate questions that are not specific enough to hint an answer, or simply nonsense questions. For instance, in (Heilman and Smith 2010a)'s work on ranking questions, most unacceptable questions are due to two factors: "do not make sense" (e.g. "`Who was the investment?`") and vagueness (e.g. "`What do modern cities also have?`" from "`modern cities also have many problems`"). Thus generating questions that are more clear by themselves is important to human-computer communication. Selecting a better question word can be one possible solution.

In the current stage of question generation, question words are confined to the following:

- WH questions, i.e., who, when, where, what, which, why, what if.
- HOW questions, i.e., how many, how much, how.
- YES/NO questions.

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Arguably "what" questions are a main contributing factor to question vagueness, as the question word itself does not provide any extra information about the answer. Consider the following input sentence and generated questions (adapted from a real-world experiment result from (Yao et al. 2011)):

(1) `A sword is made of metal.`

Questions on "`a sword`":

(a) `What is made of metal?`

(b) `What weapon is made of metal?`

Questions on "`metal`":

(c) `What is a sword made of?`

(d) `What material is a sword made of?`

(e) `What music type is a sword made of?`

Questions (a) to (e) reveal two major problems of QG:

1. Questions are not informative (in a sense that people do not know what the question asks about). Compared to (a), (b) is a much better question as it provides the scope of the answer. Question (c) would have been as bad as (a) if the words "`made of`" did not provide any extra semantic clue on the answer.
2. Questions are not precise (in a sense that the question word is wrong). Comparing (c) and (d), it is good to recognize "`metal`" as possibly a type of music but the generator failed to disambiguate the question word given the context.

Problem 2 is a natural consequence of problem 1: once computers have the power to ask more specific questions, they have to choose what additional information is appropriate to provide. In this paper we attack these two problems by disambiguating the hypernym relation between the target words (such as "`metal`") and the question words (such as "`what material`") given the context (such as "`sword`"). Additional lexical semantic resources, specifically WordNet and Wikipedia, are used as hypernym inventories. We implemented some simple disambiguation methods based on pointwise mutual information and co-occurrence. By expanding an existing syntax-based question generation system with these disambiguated hypernyms, our new question generation system obtained a much better coverage on generating which questions.

Section 2 describes the procedure and issues of finding hypernyms to supplement question words. Section 3

proposes 4 different scoring measures for disambiguating hypernyms. We introduce the question generation system briefly in Section 4. The actual evaluation results on hypernym accuracy and question quality are reported in Section 5 and Section 6. Finally we refer to related work in Section 7 and conclude in Section 8.

2 Finding Hypernyms

We rely on using existing named entity recognizers (NERs) to find hypernyms for target words. However, most NERs follow the format of the Named Entity Tasks of the Message Understanding Conferences (MUC). Thus they only produce a limited tag set, i.e., PERSON, LOCATION, ORGANIZATION and MISC. With the aid of regular expressions, some tools can also produce tags for date, phone number, email address, etc. However, these types are somehow still too general. For instance, both the World Health Organization and universities are recognized as ORGANIZATION by these taggers. While it is common practice to regard WHO as an organization, it is quite uncommon to use "`which organization`" to refer to a school. Questions like "`which organization did Prince William attend?`" or even "`what did Prince William attend?`" might be very misleading to people. Moreover, even though location related questions are generally easier to answer, a good question can still further provide extra information on the type of location, e.g., a continent, a country, a city, etc.

Thus we decided to employ a named entity recognition approach based on gazetteer matching. Note that there is plenty of previous work on finding hypernym-hyponym relations between words (See Related Work in Section 7). However, we do not apply these approaches of trying to discover our own hypernyms but just instead simply use predefined ontologies. The main reason is that no approaches guarantee a precise generation of hypernyms given only a hyponym without context. Thus a following disambiguation model is necessary to put context into consideration. The task of context-sensitive question type disambiguation boils down to the traditional task of word sense disambiguation (WSD). The difference is that in WSD tasks a dictionary has to be used to tell different senses apart, while in question type disambiguation we only need to select the most similar hypernym of a target word given a context. We employ two dictionary-like databases: WordNet (Fellbaum 1998) and WikiNet (Nastase et al. 2010).

WordNet groups words into synonym sets (or *synsets*) and builds a hierarchical structure among the synsets. Different level of synsets connects by hypernym-hyponym relations. WordNet version 3.0 contains about 118 thousand nouns. Among them, 90% is monosemous thus has only one synset as its hypernym. For instance, Scotland is a type of country, state or land. The other 10% is polysemous. Usually they are commonly used nouns and have multiple hypernym relations with multiple synsets. In this case context needs to be used to help select the best hypernym.

One shortcoming of WordNet is that it contains very few proper nouns that refer to historical events, books, movies, etc. For instance, it does not recognize "Los Angeles Times" as a newspaper and "Forrest Gump" as a person or movie. Thus we use WikiNet as a complement.

WikiNet is a large multi-lingual *concept* network built on Wikipedia. Concept is an abstract layer for Wikipedia articles and categories. It usually has multi-lingual lexical instances since most Wikipedia articles have translations from different languages. All concepts assemble a network following the original hierarchy of Wikipedia categories. These categories, along with the syntactic parses of articles, help extract relations among concepts. For instance, a *directed_by* relation can be extracted from most articles about movies (e.g. Forrest Gump was *directed_by* Robert Zemeckis). Also, the article and its parent category form an *isa* relation (e.g. Forrest Gump *isa* 1994 film, American film, etc). (Nastase and Strube 2008) gives a detailed description of the extraction.

Currently WikiNet has around 3.7 million concepts. We are mainly interested in the 10.2 million *isa* relations to find out hypernyms for multi-word terms.

3 Disambiguating Hypernyms

Given the large amount of named entity types (or hypernyms) we use in this task, it is impractical to apply a supervised approach for disambiguation. Not only because it is costly, but also that it is confined to the domain of training data. A free and open-domain alternative is the web. Motivated by (Church and Hanks 1990), we use the Pointwise Mutual Information (PMI) between the answer and the question word to select the preferred hypernym*:

$$\begin{aligned}\text{hypernym}^* &= \text{argmax}_i I(\text{hypernym}_i, \text{answer}) \\ &= \text{argmax}_i \log_2 \frac{P(\text{hypernym}_i, \text{answer})}{P(\text{hypernym}_i)P(\text{answer})} \\ &= \text{argmax}_i \frac{P(\text{hypernym}_i, \text{answer})}{P(\text{hypernym}_i)}\end{aligned}$$

In practice, we follow (Turney 2001) and use the search engine count as a simple substitution for the probability. Also, the original context has to be considered for disambiguation. Here we propose four different scores to select the best hypernym.

Score 1 is a PMI-based measure between hypernym and the answer given the context. $c(\cdot)$ is the count of documents returned by Microsoft Bing:

$$s_1 = \frac{c(\text{hypernym}_i, \text{answer}, \text{context})}{c(\text{hypernym}_i, \text{context})}$$

Score 2 uses Microsoft Bing's NEAR operator to confine that the hypernym is within 10 word window of the answer (inspired by score 4 of (Turney 2001)):

$$s_2 = \frac{c(\text{hypernym}_i \; \textit{NEAR} \; \text{answer}, \text{context})}{c(\text{hypernym}_i, \text{context})}$$

Score 3 simply uses the counts where the hypernym and context co-occur. It is mostly an engineer tweak as that oftentimes the counts provided by search engines are not reliable

thus in practice PMI-based scoring might not work as well as simple counting:

$$s_3 = c(\text{hypernym}_i, \text{context})$$

Score 4 counts the co-occurrences of the hypernym, the answer and the context as the answer itself might provide some clue for the hypernym.

$$s_4 = c(\text{hypernym}_i, \text{answer}, \text{context})$$

Note that in score 1 and 2 the best hypernym should have the lowest score value while in score 3 and 4 the best hypernym should have the highest score value.

4 Question Generation System

We used Michael Heilman's open source question generation system, Question Transducer (Heilman and Smith 2009), as our baseline system. It employs a pipeline of three stages: transforming declarative sentences, creating questions and ranking questions.

Stage 1 mainly extracts simple sentences from long sentences (Heilman and Smith 2010b). Question Transducer targets on asking questions about facts thus it extracts simplified factual statement. A series of syntactic patterns is predefined to match the trees of input sentences. If matched, then a new tree is extracted and possibly reordered to make a new sentence. The tree matching and operations are performed by the Stanford Tregex and Tsurgeon software (Levy and Andrew 2006). Examples of this simplification includes extraction from non-restrictive appositives, non-restrictive relative clauses etc.

Stage 2 mainly contains three steps. First, phrases that cannot undergo WH-movement to generate a question are marked. For instance, it is ungrammatical to generate the question "`who did John meet and Mary?`" from "`John met Bob and Mary`". Thus phrases under conjunctions, such as "`Bob`" and "`Mary`", are marked unmovable. Second, the possible answer phrases are identified with a SuperSense tagger (Ciaramita and Altun 2006), which jointly disambiguates noun phrases and tags them with proper semantic classes (such as PERSON, LOCATION and ORGANIZATION). Finally, the answer phrases are replaced by corresponding question words and the whole sentence is re-ordered and processed to form a question.

Stage 3 ranks questions (Heilman and Smith 2010a) with a linear regression model. The feature sets cover the length, question words, language model scores, grammaticality etc. Evaluation shows that about 4 questions out of the top 10 are acceptable by native speakers.

Question Transducer is able to generate who, what, where, when, whose, and how many questions. We supplemented it with which questions to provide more information for the question word. In order to do so, we added two more taggers based on WordNet and WikiNet besides the original SuperSense tagger. Since WordNet is produced by professional lexicographers and contains more accurate data than WikiNet, we decided not to use WikiNet to find hypernyms if a term can be found in WordNet. However, because WordNet contains mostly single word terms, the WikiNet tags are mostly fired for multi-word terms. The whole procedure still follows the three-step pipeline in Question Transducer. But question word selection is different:

1. Get a list of candidate answer phrases in the sentence and tag each of them with:
   (a) SuperSense tagger, to return its semantic class
   (b) WordNet tagger, to return its hypernyms
   (c) WikiNet tagger, if it cannot be found in WordNet
2. Use the best scoring function defined in Section 3 to rank all the hypernyms given context.
3. Question word selection:
   (a) For SuperSense tags, follow the original question word generation procedure, e.g., LOCATION generates where questions.
   (b) For the top-ranked hypernym, generate corresponding question words. For instance, if the hypernym for "metal" is "material", then the question word would be "which material".

The final questions still go through the ranking step. We did not add additional features based on hypernym source and disambiguate scores. The question ranker works as it is.

5 Evaluation On Hypernyms

Before we make the question generation system use the hypernyms, we need to know how good those hypernyms are. Thus we performed a manual check. We used the development data released by the the First Question Generation Shared Task Evaluation Challenge (QGSTEC2010, (Rus et al. 2010)). The data contains 24 sentences from Wikipedia, 31 from OpenLearn and 26 from Yahoo!Answers (81 in total), along with some manually written questions (180 in total). In this evaluation, we only calculated the precision of the hypernym taggers and did not make use of the written questions.

Table 1 shows the result. Out of all 81 sentences, the WordNet and WikiNet tagger found hypernyms for 231 terms (including single word terms and multi-word terms, without overlap). Since each term can have a list of hypernyms instead of only one, we manually checked whether these hypernyms contain the most appropriate one. Out of the 231 terms, the taggers provided at least one correct hypernym for 138 terms, counting 59.7% of all, thus the *oracle* column. However, it is not necessary that the disambiguator always select the best one from the candidate hypernym list. The co-occurrence based scores (3 and 4) outperformed the mutual information based scores (1 and 2). The best measure score 4, counting the co-occurrence of the hypernym, answer phrase and context, had 94 terms right out of all 138 terms that could be made right. In other words, if the hypernym tagger were able to provide a list that contains at least one correct hypernym, then the score 4 measure would select it 68.1% of the time. As a reference, the hypernym taggers output 1579 hypernyms for the 231 terms, or 6.8 hypernyms per term.

	all	oracle	score 1	score 2	score 3	score 4
number	231	138	77	82	88	94
number/all	100%	59.7%	33.3%	35.5%	38.1%	40.1%
number/oracle	-	100%	55.8%	59.4%	63.8%	68.1%

Table 1: The number and percentage of terms that have the right hypernym selected. The oracle column is the count of items that has a right hypernym in the hypernym list produced by WordNet and WikiNet. number/all represents the percentage of right hypernyms found for all terms. number/oracle represents the percentage of right hypernyms out of all found hypernyms that could have been right.

It is worth pointing out that when performing the oracle evaluation, the criterion of being right was much stricter than simply checking whether the hypernym is actually a real hypernym for the answer. Instead, we used a substitution method to also take language naturalness and context into account. Recall that the final objective is to ask *natural* which questions. Thus we substituted all the answer phrases with its corresponding "`which + hypernym`" question words to make a question, then judged whether the question was natural. Any hypernyms that sound awkward in "`which + hypernym`" question words were regarded wrong. For instance, one of the hypernyms for "`hair`" is "`material`". But in most sentences asking a "`which material`" question is misleading, e.g., "`what material do you have on your head`?". A correct hypernym would be "`body covering`".

Given the strict checking criterion, we were not surprised by the 40.1% overall precision (after disambiguation) for the hypernyms.

6 Evaluation on Questions

In this section we first introduce the evaluation criterion, then report the question coverage on the test set. Fianlly we show and discuss the question quality given the evaluation criterion.

6.1 Criterion

The final evaluation on generated questions was performed on the test set of QGSTEC2010. Very similar as the development data, the test set contains 90 input sentences in all from Wikipedia, OpenLearn and Yahoo!Answers. For each sentence a question generation system is required to generate 1 to 4 types of questions, with each type two different questions. In total there are 360 questions required.

Following QGSTEC2010, the evaluation criteria measure different aspects of question qualities. The best score is always 1 while the worse score depends on the specific criterion with an incremental size of 1. The following lists all the criteria with its grading range listed in the parenthesis.

- Relevance (1-4). Questions should be relevant to the input sentence.
- Question type (1-2). Questions should be of the specified target question type.
- Syntactic correctness and fluency (1-4). The syntactic correctness is rated to ensure systems can generate sensible output.
- Ambiguity (1-3). The question should make sense when asked more or less out of the blue.
- Variety (1-3). Pairs of questions in answer to a single input are evaluated on how different they are from each other.

6.2 Question Coverage

Out of all required question types (who, what, where, when, why, which, how many and `yes/no`), the original Question Transducer does not have why and which questions pre-encoded. Expanding it with hypernyms makes Question Transducer able to generate which questions and it is our main question type of evaluation. We ignore why questions in this task as it is not relevant to question word selection. Eliminating why questions reduces the total number of input sentences from 90 to 86 (as in 4 sentences only why questions are asked) and the total number of required questions from 360 to 330.

One other substantial change we have made to Question Transducer is that we enabled generating questions from unmovable phrases without moving the question words. Recall that the original system disables some syntactic movement to prevent ungrammatical question word fronting. For instance, the question "`who did John meet and Mary?`" is prohibited from "`John met Bob and Mary`". However, the evaluators are sometimes very tolerant of these questions if the question word is not fronted: "`John met whom and Mary?`". This helps increase recall without hurting score too much. For instance, when the evaluation result came back, we found out that the two raters gave good scores on the following input:

(2) `Among many others, the UK, Spain, Italy and France are unitary states, while Germany, Canada, the United States of America, Switzerland and India are federal states.`

Two generated which questions:

(a) `The UK, Spain, Italy and which European country are unitary states?`

(b) `The UK, Spain, Italy and which European nation are unitary states?`

Average score from two raters:

relevance=1 questionType=1 correctness=1 ambiguity=1.5 variety=2

	who	what	where	when	how many	yes/no	which	total
test set	30	116	30	36	46	28	44	330
QTorig	28	100	8	26	16	24	0	202
QThyp	28	100	8	26	16	24	38	240

Table 2: The number of questions asked in the test set and actually generated by the original system, QTorig and the hypernym-expanded system, QThyp. The only difference between the two systems is that QThyp generates 38 more which questions than QTorig. All other questions are identical.

Out of the 330 questions to be generated, 44 of them are which questions. The original system, QTorig, failed to generate any of those as which questions were not encoded in the system. The hypernym-expanded system, QThyp, was able to generate 38 of them. Table 2 lists the number of questions generated before and after using hypernyms.

One can play a trick in the original system to convert all `who/where/when` questions to `which person/location/time` questions and thus generate some of the `which` questions. However, we argue that these questions are in fact just a variant of the `who/where/when` questions and do not show the system's ability to choose from different question types given the context. Thus we did not play this trick for evaluation. Interested readers who want to know the quality of these types of `which` questions can refer to the scores for `who/where/when` questions as they are basically the same questions under different question words.

6.3 Question Quality

We asked two native English speakers to rate the 240 generated questions in total. Following the convention of QGSTEC2010[1], we report evaluation score per question type. These scores are based only on generated questions, without enforcing any penalty on missing questions. Table 4 shows the result.

On average, which questions receive worse score than other question types. The low relevance and ambiguity scores indicate that wrong types of hypernyms make the question more irrelevant to the original input sentence and confuse the human raters. A manual check shows that the better score in variance than where, when, how many and yes/no questions is a result of multiple choices of hypernyms, rather than variant questions of different answer phrases. The low human rater agreement and Cohen's Kappa in relevance, correctness and ambiguity reflect the subjective nature of these criteria, consistent with the result of QGSTEC2010[2].

To best investigate what factor contributes to bad score of which questions, we did a manual check on the accuracy of question words, shown below in Table 3. Among all the question types that can be determined by a named entity recognizer, when, how many, who and where questions come from the SuperSense tagger, which was trained in a supervised fashion, thus achieving an accuracy between 75% and 90%. WordNet and WikiNet produced hypernyms to construct which questions that were disambiguated in an unsupervised fashion, only achieving an accuracy of 55.3%.

	correct	all	accuracy
when	33	26	88.5%
how many	14	16	87.5%
who	23	28	82.1%
where	6	8	75.0%
which	21	38	55.3%

Table 3: The accuracy of question words in the test set.

We further evaluated the 21 out of 38 which questions that have the correct hypernyms, shown in the which* column of Table 4. This time which* questions receive better scores than most other question types, showing the correctness of hypernyms is a deciding factor to the quality of which questions.

7 Related Work

Along the line of performing syntactic transformation for question generation (Heilman and Smith 2009; Wyse and Piwek 2009), there are also works based on templates (Mostow and Chen 2009; Chen, Aist, and Mostow 2009) and semantic transformation (Schwartz, Aikawa, and Pahud 2004; Sag and Flickinger 2008; Yao 2010). Most of these systems confine the questions words with traditional named entity recognizer output (person, location and organization) or template definition. Thus selecting a proper question word given the context has not been attempted in previous work.

A key step towards selecting better question words is to find appropriate hypernyms. There is plenty of research on finding hypernym-hyponym relations between words (a few of them: (Hearst 1992; Snow, Jurafsky, and Ng 2005; Garera and Yarowsky 2006; Kozareva, Riloff, and Hovy 2008)). However, not much work has been done in disambiguating hypernyms with context. (Turney 2001)'s idea on

[1] It is not fair to compare the result of Question Transducer with that of participating systems in QGSTEC2010 as Question Transducer in this paper was used as it is, and not optimized for an evaluation challenge.

[2] http://www.questiongeneration.org/QGSTEC2010/

	who	what	where	when	how many	yes/no	which	which*	average	agreement	Kappa
relevance	1.50	1.51	1.94	1.21	1.47	1.21	1.89	1.38	1.52	0.54	0.25
ques. type	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
correctness	2.14	1.93	1.63	2.71	1.91	1.44	2.13	1.63	2.01	0.57	0.43
ambiguity	1.70	1.78	1.63	1.40	1.56	1.19	1.92	1.35	1.67	0.61	0.44
variety	1.57	1.22	2.63	2.31	2.31	1.88	1.71	1.80	1.64	0.84	0.76

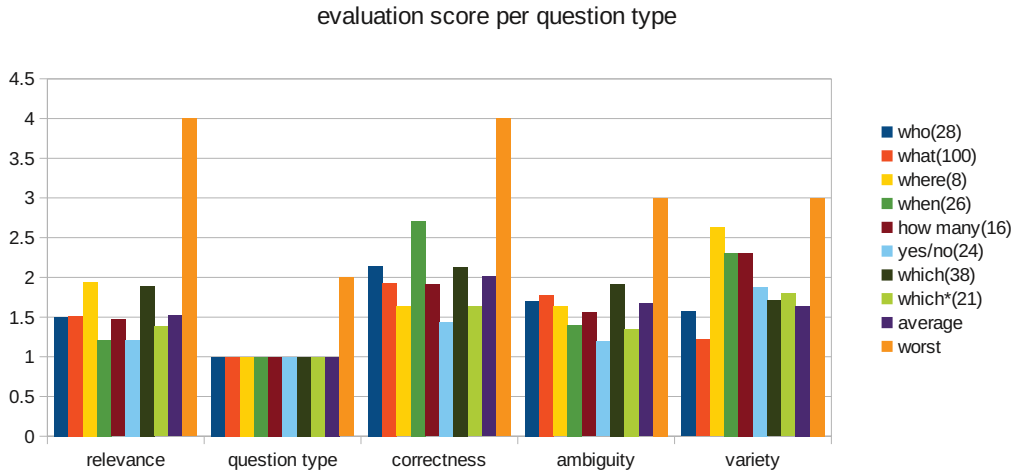

Table 4: Evaluation score per question type shown numerically in a table and graphically in a chart. Lower scores are better while the worst scores are indicated in the chart. The number in the parenthesis of the chart labels indicates how many questions were actually generated, e.g., who(28) shows th 28 who questions were generated and evaluated. The which* column shows evaluation result on those which questions (21) that have selected the correct hypernym out of all which questions (38).

selecting synonyms inspired this paper. Using a search engine to approximate probability counts can be problematic in some sense (Kilgarriff 2007), but there is also research reporting positive result (Keller, Lapata, and Ourioupina 2002). In this paper's particular task setting, to quickly employ an unsupervised disambiguation method based on a large amount of data, search engines can be a solution. Other approaches use more complicated methods, such as vector space models (Turney and Pantel 2010).

8 Conclusion

We attacked the problem of how to ask more specific questions by supplementing question words with the hypernyms for answer phrases, in the hope that more informative question words leave more hint to the answer, thus producing less ambiguous questions. We used WordNet and WikiNet as hypernym inventories, which provide a list of hypernyms that contains a correct hypernym 59.7% of the time. Several simple disambiguation methods based on pointwise mutual information and co-occurrences are proposed, compared and evaluated, with the best one being able to find a correct hypernym 68.1% of the time. Furthermore, an existing question generation system was expanded by utilizing disambiguated hypernyms as more informative question words. Evaluation results show that the accuracy of hypernym disambiguation method is the deciding factor for the quality of produced questions, as most low scores are due to bad question words. Future work includes expanding question words beyond hypernyms, such as using holonym and meronym relations. Also, more accurate disambiguation methods for the question words need to be developed for better question qualities.

9 Acknowledgements

We thank Shane Bergsma and Benjamin Van Durme for helpful discussions. We are also grateful to the careful work of two raters.

Appendix

Here we list some which questions that were actually generated given the test set to give readers some sense of the quality of these questions.

1. Input:
   `Rashid (1998) points out that in markets where entry barriers are low, quality has the greatest tendency to vary.`
   Questions:
   (a) `Rashid 1998 points out ..., which social rank has the greatest tendency to vary?`
   (b) `Rashid 1998 points out ..., which social station has the greatest tendency to vary?`

2. Input:
   Designer Ian Callum, originally from Dumfries in Scotland, studied at the Glasgow School of Art and at *the Royal College* of Art in *London*.
   Questions:
   (a) *Which national capital* did designer Ian Callum study at the Glasgow School of Art and at the Royal College of Art in?
   (b) Designer Ian Callum studied at the Glasgow School of Art and at *which professional association* of Art in London?

3. Input:
   The Tigris unites with *the Euphrates* near Basra, and from this junction to the Persian Gulf the mass of moving water is known as the Shatt-al-Arab.
   Questions:
   (a) The Tigris unites with *which river* near Basra?
   (b) *Which river* unites with the Euphrates near Basra?

4. Input:
   According to the Biblical book of Daniel, at a young age *Daniel* was carried off to Babylon where he was trained in the service of the court under the authority of Ashpenaz.
   Questions:
   (a) *Which jurist* was carried off to Babylon where he was trained in the service of the court under the authority of Ashpenaz at a young age?
   (b) *Which prophet* was carried off to Babylon where he was trained in the service of the court under the authority of Ashpenaz at a young age?

5. Input:
   Furthermore, Softbank BB's president, *Masayoshi Son*, announced that he and other senior executives would take a 50 per cent pay cut for the next six months.
   Questions:
   (a) *Which alumnus* was Softbank BB's president?
   (b) *Which person* was Softbank BB's president?

References

Chen, W.; Aist, G.; and Mostow, J. 2009. Generating Questions Automatically from Informational Text. In *Proceedings of the 2nd Workshop on Question Generation In Craig, S.D. & Dicheva, S. (Eds.) (2009) AIED 2009: 14th International Conference on Artificial Intelligence in Education: Workshops Proceedings*.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.* 16:22–29.

Ciaramita, M., and Altun, Y. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 594–602. Stroudsburg, PA, USA: Association for Computational Linguistics.

Fellbaum, C. 1998. *WordNet: An Electronical Lexical Database*.

Garera, N., and Yarowsky, D. 2006. Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, 37–44.

Hearst, M. A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *In Proceedings of the 14th International Conference on Computational Linguistics*, 539–545.

Heilman, M., and Smith, N. A. 2009. Question Generation via Overgenerating Transformations and Ranking. Technical report, Language Technologies Institute, Carnegie Mellon University Technical Report CMU-LTI-09-013.

Heilman, M., and Smith, N. A. 2010a. Good Question! Statistical Ranking for Question Generation. In *Proc. of NAACL/HLT*.

Heilman, M., and Smith, N. A. 2010b. Extracting Simplified Statements for Factual Question Generation. In *Proceedings of the 3rd Workshop on Question Generation.*

Keller, F.; Lapata, M.; and Ourioupina, O. 2002. Using the web to overcome data sparseness. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, 230–237. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kilgarriff, A. 2007. Googleology is Bad Science. *Comput. Linguist.* 33(1):147–151.

Kozareva, Z.; Riloff, E.; and Hovy, E. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *Proceedings of ACL-08: HLT*, 1048–1056. Columbus, Ohio: Association for Computational Linguistics.

Levy, R., and Andrew, G. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Mostow, J., and Chen, W. 2009. Generating Instruction Automatically for the Reading Strategy of Self-Questioning. In *Proceeding of the 2009 conference on Artificial Intelligence in Education*, 465–472. Amsterdam, The Netherlands: IOS Press.

Nastase, V., and Strube, M. 2008. Decoding wikipedia categories for knowledge acquisition. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, 1219–1224. AAAI Press.

Nastase, V.; Strube, M.; Boerschinger, B.; Zirn, C.; and Elghafari, A. 2010. WikiNet: A Very Large Scale Multi-Lingual Concept Network. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Rus, V.; Wyse, B.; Piwek, P.; Lintean, M.; Stoyanchev, S.; and Moldovan, C. 2010. Overview of The First Question Generation Shared Task Evaluation Challenge. In Boyer, K. E., and Piwek, P., eds., *Proceedings of the Third Workshop on Question Generation*.

Sag, I. A., and Flickinger, D. 2008. Generating Questions with Deep Reversible Grammars. In *Proceedings of the First Workshop on the Question Generation Shared Task and Evaluation Challenge*.

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

Snow, R.; Jurafsky, D.; and Ng, A. Y. 2005. Learning Syntactic Patterns for Automatic Hypernym Discovery. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press. 1297–1304.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.* 37:141–188.

Turney, P. D. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, 491–502. London, UK: Springer-Verlag.

Wyse, B., and Piwek, P. 2009. Generating Questions from OpenLearn study units. In *Proceedings of the 2nd Workshop on Question Generation In Craig, S.D. & Dicheva, S. (Eds.) (2009) AIED 2009: 14th International Conference on Artificial Intelligence in Education: Workshops Proceedings*.

Yao, X.; Tosch, E.; Chen, G.; Nouri, E.; Artstein, R.; Leuski, A.; Sagae, K.; and Traum, D. 2011. Creating Conversational Characters Using Question Generation Tool. *Dialogue and Discourse: Special Issue on Question Generation*.

Yao, X. 2010. Question Generation with Minimal Recursion Semantics. Master's thesis, Saarland University & University of Groningen.