Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding

Delphine Bernhard and **Iryna Gurevych**
Ubiquitous Knowledge Processing (UKP) Lab
Computer Science Department
Technische Universität Darmstadt, Hochschulstraße 10
D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de/

Abstract

Monolingual translation probabilities have recently been introduced in retrieval models to solve the *lexical gap* problem. They can be obtained by training statistical translation models on parallel monolingual corpora, such as question-answer pairs, where answers act as the "source" language and questions as the "target" language. In this paper, we propose to use as a parallel training dataset the definitions and glosses provided for the same term by different lexical semantic resources. We compare monolingual translation models built from lexical semantic resources with two other kinds of datasets: manually-tagged question reformulations and question-answer pairs. We also show that the monolingual translation probabilities obtained (i) are comparable to traditional semantic relatedness measures and (ii) significantly improve the results over the query likelihood and the vector-space model for answer finding.

1 Introduction

The *lexical gap* (or *lexical chasm*) often observed between queries and documents or questions and answers is a pervasive problem both in Information Retrieval (IR) and Question Answering (QA). This problem arises from alternative ways of conveying the same information, due to synonymy or paraphrasing, and is especially severe for retrieval over shorter documents, such as sentence retrieval or question retrieval in Question & Answer archives. Several solutions to this problem have been proposed including query expansion (Riezler et al., 2007; Fang, 2008), query reformulation or paraphrasing (Hermjakob et al., 2002; Tomuro, 2003; Zukerman and Raskutti, 2002) and semantic information retrieval (Müller et al., 2007).

Berger and Lafferty (1999) have formulated a further solution to the lexical gap problem consisting in integrating monolingual statistical translation models in the retrieval process. Monolingual translation models encode statistical word associations which are trained on parallel monolingual corpora. The major drawback of this approach lies in the limited availability of truly parallel monolingual corpora. In practice, training data for translation-based retrieval often consist in question-answer pairs, usually extracted from the evaluation corpus itself (Riezler et al., 2007; Xue et al., 2008; Lee et al., 2008). While collection-specific translation models effectively encode statistical word associations for the target document collection, it also introduces a bias in the evaluation and makes it difficult to assess the quality of the translation model per se, independently from a specific task and document collection.

In this paper, we propose new kinds of datasets for training domain-independent monolingual translation models. We use the definitions and glosses provided for the same term by different lexical semantic resources to automatically train the translation models. This approach has been very recently made possible by the emergence of new kinds of lexical semantic and encyclopedic resources such as Wikipedia and Wiktionary. These resources are freely available, up-to-date and have a broad coverage and good quality. Thanks to the combination of several resources, it is possible to obtain monolingual parallel corpora which are large enough to train domain-independent translation models. In addition, we collected question-answer pairs and manually-tagged question reformulations from a social Q&A site. We use these datasets to build further translation models.

Translation-based retrieval models have been

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 728–736,
Suntec, Singapore, 2-7 August 2009.

widely used in practice by the IR and QA community. However, the quality of the semantic information encoded in the translation tables has never been assessed intrinsically. To do so, we compare translation probabilities with concept vector based semantic relatedness measures with respect to human relatedness rankings for reference word pairs. This study provides empirical evidence for the high quality of the semantic information encoded in statistical word translation tables. We then use the translation models in an answer finding task based on a new question-answer dataset which is totally independent from the resources used for training the translation models. This extrinsic evaluation shows that our translation models significantly improve the results over the query likelihood and the vector-space model.

The remainder of the paper is organised as follows. Section 2 discusses related work on semantic relatedness and statistical translation models for retrieval. Section 3 presents the monolingual parallel datasets we used for obtaining monolingual translation probabilities. Semantic relatedness experiments are detailed in Section 4. Section 5 presents answer finding experiments. Finally, we conclude in Section 6.

2 Related Work

2.1 Statistical Translation Models for Retrieval

Statistical translation models for retrieval have first been introduced by Berger and Lafferty (1999). These models attempt to address synonymy and polysemy problems by encoding statistical word associations trained on monolingual parallel corpora. This method offers several advantages. First, it bases upon a sound mathematical formulation of the retrieval model. Second, it is not as computationally expensive as other semantic retrieval models, since it only relies on a word translation table which can easily be computed before retrieval. The main drawback lies in the availability of suitable training data for the translation probabilities. Berger and Lafferty (1999) initially built synthetic training data consisting of queries automatically generated from documents. Berger et al. (2000) proposed to train translation models on question-answer pairs taken from Usenet FAQs and call-center dialogues, with answers corresponding to the "source" language and questions to the "target" language.

Subsequent work in this area often used similar kinds of training data such as question-answer pairs from Yahoo! Answers (Lee et al., 2008) or from the Wondir site (Xue et al., 2008). Lee et al. (2008) tried to further improve translation models based on question-answer pairs by selecting the most important terms to build compact translation models.

Other kinds of training data have also been proposed. Jeon et al. (2005) automatically clustered semantically similar questions based on their answers. Murdock and Croft (2005) created a first parallel corpus of synonym pairs extracted from WordNet, and an additional parallel corpus of English words translating to the same Arabic term in a parallel English-Arabic corpus.

Similar work has also been performed in the area of query expansion using training data consisting of FAQ pages (Riezler et al., 2007) or queries and clicked snippets from query logs (Riezler et al., 2008).

All in all, translation models have been shown to significantly improve the retrieval results over traditional baselines for document retrieval (Berger and Lafferty, 1999), question retrieval in Question & Answer archives (Jeon et al., 2005; Lee et al., 2008; Xue et al., 2008) and for sentence retrieval (Murdock and Croft, 2005).

Many of the approaches previously described have used parallel data extracted from the retrieval corpus itself. The translation models obtained are therefore domain and collection-specific, which introduces a bias in the evaluation and makes it difficult to assess to what extent the translation model may be re-used for other tasks and document collections. We henceforth propose a new approach for building monolingual translation models relying on domain-independent lexical semantic resources. Moreover, we extensively compare the results obtained by these models with models obtained from a different type of dataset, namely Question & Answer archives.

2.2 Semantic Relatedness

The rationale behind translation-based retrieval models is that monolingual translation probabilities encode some form of semantic knowledge. The semantic similarity and relatedness of words has traditionally been assessed through corpus-based and knowledge-based measures. Corpus-based measures include Hyperspace Analogue to

Language (HAL) (Lund and Burgess, 1996) and Latent Semantic Analysis (LSA) (Landauer et al., 1998). Knowledge-based measures rely on lexical semantic resources such as WordNet and comprise path length based measures (Rada et al., 1989) and concept vector based measures (Qiu and Frei, 1993). These measures have recently also been applied to new collaboratively constructed resources such as Wikipedia (Zesch et al., 2007) and Wiktionary (Zesch et al., 2008), with good results.

While classical measures of semantic relatedness have been extensively studied and compared, based on comparisons with human relatedness judgements or word-choice problems, there is no comparable intrinsic study of the relatedness measures obtained through word translation probabilities. In this study, we use the correlation with human rankings for reference word pairs to investigate how word translation probabilities compare with traditional semantic relatedness measures. To our knowledge, this is the first time that word-to-word translation probabilities are used for ranking word-pairs with respect to their semantic relatedness.

3 Parallel Datasets

In order to obtain parallel training data for the translation models, we collected three different datasets: manually-tagged question reformulations and question-answer pairs from the WikiAnswers social Q&A site (Section 3.1), and glosses from WordNet, Wiktionary, Wikipedia and Simple Wikipedia (Section 3.2).

3.1 Social Q&A Sites

Social Q&A sites, such as Yahoo! Answers and AnswerBag, provide portals where users can ask their own questions as well as answer questions from other users.

For our experiments we collected a dataset of questions and answers, as well as question reformulations, from the WikiAnswers[1] (WA) web site. WikiAnswers is a social Q&A site similar to Yahoo! Answers and AnswerBag. The main originality of WikiAnswers is that users might manually tag question reformulations in order to prevent the duplication of answers to questions asking the same thing in a different way. When a user enters a question that is not already part of the question repository, the web site displays a list of already existing questions similar to the one just asked by the user. The user may then freely select the question which paraphrases her question, if available. The question reformulations thus labelled by the users are stored in order to retrieve the same answer when a given question reformulation is asked again.

We collected question-answer pairs and question reformulations from the WikiAnswers site. The resulting dataset contains 480,190 questions with answers.[2] We use this dataset in order to train two different translation models:

Question-Answer Pairs (WAQA) In this setting, question-answer pairs are considered as a parallel corpus. Two different forms of combinations are possible: (Q,A), where questions act as source and answers as target, and (A,Q), where answers act as source and questions as target. Recent work by Xue et al. (2008) has shown that the best results are obtained by pooling the question-answer pairs $\{(q,a)_1,...,(q,a)_n\}$ and the answer-question pairs $\{(a,q)_1,...,(a,q)_n\}$ for training, so that we obtain the following parallel corpus: $\{(q,a)_1,...,(q,a)_n\} \cup \{(a,q)_1,...,(a,q)_n\}$. Overall, this corpus contains 1,227,362 parallel pairs and will be referred to as **WAQA** (WikiAnswers Question-Answers) in the rest of the paper.

Question Reformulations (WAQ) In this setting, question and question reformulation pairs are considered as a parallel corpus, e.g. *'How long do polar bears live?'* and *'What is the polar bear lifespan?'*. For a given user question q_1, we retrieve its stored reformulations from the WikiAnswers dataset; $q_{11}, q_{12}, ...$ The original question and reformulations are subsequently combined and pooled to obtain a parallel corpus of question reformulation pairs: $\{(q_1,q_{11}),(q_1,q_{12}),...,(q_n,q_{nm})\} \cup \{(q_{11},q_1),(q_{12},q_1),...,(q_{nm},q_n)\}$. This corpus contains 4,379,620 parallel pairs and will be referred to as **WAQ** (WikiAnswers Questions) in the rest of the paper.

3.2 Lexical Semantic Resources

Glosses and definitions for the same lexeme in different lexical semantic and encyclopedic resources can actually be considered as near-paraphrases, since they define the same terms and hence have

[1] http://wiki.answers.com/

[2] A question may have more than one answer.

gem				moon			
WAQ	WAQA	LSR	ALL_{Pool}	WAQ	WAQA	LSR	ALL_{Pool}
gem	explorer	gem	gem	moon	moon	moon	moon
95	ford	diamonds	xlt	land	earth	lunar	land
xlt	gem	gemstone	95	foot	lunar	sun	earth
module	xlt	diamond	explorer	armstrong	apollo	earth	landed
stones	demand	natural	gemstone	set	landed	tides	armstrong
expedition	lists	facets	diamonds	actually	neil	moons	neil
ring	dash	rare	natural	neil	1969	phase	apollo
gemstone	center	synthetic	diamond	landed	armstrong	crescent	set
modual	play	ruby	ford	apollo	space	astronomical	foot
crystal	lights	usage	ruby	walked	surface	occurs	actually

Table 1: Sample top translations for different training data. ALL corresponds to WAQ+WAQA+LSR.

the same meaning, as shown by the following example for the lexeme "moon":

- *Wordnet (sense 1)*: the natural satellite of the Earth.
- *English Wiktionary*: The Moon, the satellite of planet Earth.
- *English Wikipedia*: The Moon (Latin: Luna) is Earth's only natural satellite and the fifth largest natural satellite in the Solar System.

We use glosses and definitions contained in the following resources to build a parallel corpus:

- WordNet (Fellbaum, 1998). We use a freely available API for WordNet (JWNL[3]) to access WordNet 3.0.
- English Wiktionary. We use the Wiktionary dump from January 11, 2009.
- English and Simple English Wikipedia. We use the Wikipedia dump from February 6, 2007 and the Simple Wikipedia dump from July 24, 2008. The Simple English Wikipedia is an English Wikipedia targeted at non-native speakers of English which uses simpler words than the English Wikipedia. Wikipedia and Simple Wikipedia articles do not directly correspond to glosses such as those found in dictionaries, we therefore considered the first paragraph in articles as a surrogate for glosses.

Given a list of 86,584 seed lexemes extracted from WordNet, we collected the glosses for each lexeme from the four English resources described above. We then built pairs of glosses by considering each possible pair of resource. Given that a lexeme might have different senses, and hence different glosses, it is possible to extract several gloss pairs for one and the same lexeme and one and the same pair of resources. It is therefore necessary to perform word sense alignment. As we do not need perfect training data, but rather large amounts of training data, we used a very simple method consisting in eliminating gloss pairs which did not at least have one lemma in common (excluding stop words and the seed lexeme itself).

The final pooled parallel corpus contains 307,136 pairs and is henceforth much smaller than the previous datasets extracted from WikiAnswers. This corpus will be referred to as **LSR**.

3.3 Translation Model Training

We used the GIZA++ SMT Toolkit[4] (Och and Ney, 2003) in order to obtain word-to-word translation probabilities from the parallel datasets described above. As is common practice in translation-based retrieval, we utilised the IBM translation model 1. The only pre-processing steps performed for all parallel datasets were tokenisation and stop word removal.[5]

3.4 Comparison of Word-to-Word Translations

Table 1 gives some examples of word-to-word translations obtained for the different parallel corpora used (the column ALL_{Pool} will be described in the next section). As evidenced by this table,

[3] http://sourceforge.net/projects/jwordnet/

[4] http://code.google.com/p/giza-pp/

[5] For stop word removal we used the list available at: http://truereader.com/manuals/onix/stopwords1.html.

the different kinds of data encode different types of information, including semantic relatedness and similarity, as well as morphological relatedness. As could be expected, the quality of the "translations" is variable and heavily dependent on the training data: the WAQ and WAQA models reveal the users' interests, while the LSR model encodes lexicographic and encyclopedic knowledge. For instance, "gem" is an acronym for "generic electronic module", which is found in Ford vehicles. Since many question-answer pairs in WA are related to cars, this very particular use of "gem" is predominant in the WAQ and WAQA translation tables.

3.5 Combination of the Datasets

In order to investigate the role played by different kinds of training data, we combined the several translation models, using the two methods described by Xue et al. (2008). The first method consists in a linear combination of the word-to-word translation probabilities *after* training:

$$\begin{aligned} P_{Lin}(w_i|w_j) &= \alpha P_{WAQ}(w_i|w_j) \\ &+ \gamma P_{WAQA}(w_i|w_j) \\ &+ \delta P_{LSR}(w_i|w_j) \end{aligned} \quad (1)$$

where $\alpha + \gamma + \delta = 1$. This approach will be labelled with the $_{\mathbf{Lin}}$ subscript.

The second method consists in pooling the training datasets, i.e. concatenating the parallel corpora, *before* training. This approach will be labelled with the $_{\mathbf{Pool}}$ subscript. Examples for word-to-word translations obtained with this type of combination can be found in the last column for each word in Table 1. The ALL_{Pool} setting corresponds to the pooling of all three parallel datasets: WAQ+WAQA+LSR.

4 Semantic Relatedness Experiments

The aim of this first experiment is to perform an intrinsic evaluation of the word translation probabilities obtained by comparing them to traditional semantic relatedness measures on the task of ranking word pairs. Human judgements of semantic relatedness can be used to evaluate how well semantic relatedness measures reflect human rankings by correlating their ranking results with Spearman's rank correlation coefficient. Several evaluation datasets are available for English, but we restrict our study to the larger dataset created by Finkelstein et al. (2002) due to the low coverage of many pairs in the word-to-word translation tables. This dataset comprises two subsets, which have been annotated by different annotators: **Fin1–153**, containing 153 word pairs, and **Fin2–200**, containing 200 word pairs.

Word-to-word translation probabilities are compared with a concept vector based measure relying on Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007), since this approach has been shown to yield very good results (Zesch et al., 2008). The method consists in representing words as a concept vector, where concepts correspond to WordNet synsets, Wikipedia article titles or Wiktionary entry names. Concept vectors for each word are derived from the textual representation available for each concept, i.e. glosses in WordNet, the full article or the first paragraph of the article in Wikipedia or the full contents of a Wiktionary entry. We refer the reader to (Gabrilovich and Markovitch, 2007; Zesch et al., 2008) for technical details on how the concept vectors are built and used to obtain semantic relatedness values.

Table 2 lists Spearman's rank correlation coefficients obtained for concept vector based measures and translation probabilities. In order to ensure a fair evaluation, we limit the comparison to the word pairs which are contained in all resources and translation tables.

Dataset	**Fin1-153**	**Fin2-200**
Word pairs used	46	42
Concept vectors		
WordNet	.26	.46
Wikipedia	.27	.03
$\text{Wikipedia}_{\text{First}}$	.30	.38
Wiktionary	.39	.58
Translation probabilities		
WAQ	.43	**.65**
WAQA	**.54**	.37
LSR	.51	.29
ALL_{Pool}	.52	.57

Table 2: Spearman's rank correlation coefficients on the Fin1-153 and Fin2-200 datasets. Best values for each dataset are in bold format. For $\text{Wikipedia}_{\text{First}}$, the concept vectors are based on the first paragraph of each article.

The first observation is that the coverage over the two evaluation datasets is rather small: only 46 pairs have been evaluated for the Fin1-153 dataset and 42 for the Fin2-200 dataset. This is mainly

due to the natural absence of many word pairs in the translation tables. Indeed, translation probabilities can only be obtained from observed parallel pairs in the training data. Concept vector based measures are more flexible in that respect since the relatedness value is based on a common representation in a concept vector space. It is therefore possible to measure relatedness for a far greater number of word pairs, as long as they share some concept vector dimensions. The second observation is that, on the restricted subset of word pairs considered, the results obtained by word-to-word translation probabilities are most of the time better than those of concept vector measures. However, the differences are not statistically significant.[6]

5 Answer Finding Experiments

5.1 Retrieval based on Translation Models

The second experiment aims at providing an extrinsic evaluation of the translation probabilities by employing them in an answer finding task. In order to perform retrieval, we use a ranking function similar to the one proposed by Xue et al. (2008), which builds upon previous work on translation-based retrieval models and tries to overcome some of their flaws:

$$P(q|D) = \prod_{w \in q} P(w|D) \quad (2)$$

$$P(w|D) = (1-\lambda)P_{mx}(w|D) + \lambda P(w|C) \quad (3)$$

$$P_{mx}(w|D) = (1-\beta)P_{ml}(w|D) + \beta \sum_{t \in D} P(w|t)P_{ml}(t|D) \quad (4)$$

where q is the query, D the document, λ the smoothing parameter for the document collection C and $P(w|t)$ is the probability of translating a document term t to the query term w.

The only difference to the original model by Xue et al. (2008) is that we use Jelinek-Mercer smoothing for equation 3 instead of Dirichlet Smoothing, as it has been done by Jeon et al. (2005). In all our experiments, β was set to 0.8 and λ to 0.5.

5.2 The Microsoft Research QA Corpus

We performed an extrinsic evaluation of monolingual word translation probabilities by integrating them in the retrieval model previously described for an answer finding task. To this aim, we used the questions and answers contained in the Microsoft Research Question Answering Corpus.[7] This corpus comprises approximately 1.4K questions collected from 10-13 year old schoolchildren, who were asked "If you could talk to an encyclopedia, what would you ask it?". The answers to the questions have been manually identified in the full text of Encarta 98 and annotated with the following relevance judgements: exact answer (1), off topic (3), on topic - off target (4), partial answer (5). In order to use this dataset for an answer finding task, we consider the annotated answers as the documents to be retrieved and use the questions as the set of test queries.

This corpus is particularly well suited to conduct experiments targeted at the lexical gap problem: only 28% of the question-answer pairs correspond to a strong match (two or more query terms in the same answer sentence), while about a half (52%) are a weak match (only one query term matched in the answer sentence) and 16 % are indirect answers which do not explicitly contain the answer but provide enough information for deducing it. Moreover, the Microsoft QA corpus is not limited to a specific topic and entirely independent from the datasets used to build our translation models.

The original corpus contained some inconsistencies due to duplicated data and non-labelled entries. After cleaning, we obtained a corpus of 1,364 questions and 9,780 answers. Table 3 gives one example of a question with different answers and relevance judgements.

We report the retrieval performance in terms of Mean Average Precision (MAP) and Mean R-Precision (R-prec), MAP being our primary evaluation metric. We consider the following relevance categories, corresponding to increasing levels of tolerance for inexact or partial answers:

- **MAP**$_1$, **R-Prec**$_1$: exact answer (1)
- **MAP**$_{1,5}$, **R-Prec**$_{1,5}$: exact answer (1) or partial answer (5)
- **MAP**$_{1,4,5}$, **R-Prec**$_{1,4,5}$: exact answer (1) or partial answer (5) or on topic - off target (4)

Similarly to the training data for translation models, the only pre-processing steps performed

[6] Fisher-Z transformation, two-tailed test with α=.05.

[7] http://research.microsoft.com/en-us/downloads/88c0021c-328a-4148-a158-a42d7331c6cf/default.aspx

Question	Why is the sun bright?
Exact answer	Star, large celestial body composed of gravitationally contained hot gases emitting electromagnetic radiation, especially light, as a result of nuclear reactions inside the star. The sun is a star.
Partial answer	Solar Energy, radiant energy produced in the sun as a result of nuclear fusion reactions (see Nuclear Energy; Sun).
On topic - off target	The sun has a magnitude of -26.7, inasmuch as it is about 10 billion times as bright as Sirius in the earth's sky.

Table 3: Example relevance judgements in the Microsoft QA corpus.

Model	**MAP**$_1$	**R-Prec**$_1$	**MAP**$_{1,5}$	**R-Prec**$_{1,5}$	**MAP**$_{1,4,5}$	**R-Prec**$_{1,4,5}$
QLM	0.2679	0.1941	0.3179	0.2963	0.3215	0.3057
Lucene	0.2705	0.2002	0.3167	0.2956	0.3192	0.3030
WAQ	0.3002	0.2149*	0.3557	0.3269	0.3583	0.3375
WAQA	0.3000	0.2211	0.3640	0.3328	0.3664	0.3405
LSR	0.3046	0.2171*	0.3666	0.3327	0.3723	0.3464
WAQ+WAQA$_{\text{Pool}}$	0.3062	0.2259	0.3685	0.3339	0.3716	0.3454
WAQ+LSR$_{\text{Pool}}$	0.3117	0.2224	0.3736	0.3399	0.3766	0.3487
WAQA+LSR$_{\text{Pool}}$	0.3135	0.2267	0.3818	0.3444	0.3840	0.3515
WAQ+WAQA+LSR$_{\text{Pool}}$	0.3152	0.2286	0.3832	0.3495	0.3848	0.3569
WAQ+WAQA+LSR$_{\text{Lin}}$	**0.3215**	**0.2343**	**0.3921**	**0.3536**	**0.3967**	**0.3673**

Table 4: Answer retrieval results. The WAQ+WAQA+LSR$_{\text{Lin}}$ results have been obtained with α=0.2 γ=0.2 and δ=0.6 (the parameter values have been determined empirically based on MAP and R-Prec). The performance gaps between the translation-based models and the baseline models are statistically significant, except for those marked with a '*' (two-tailed paired t-test, $p < 0.05$).

for this corpus were tokenisation and stop word removal. Due to the small size of the answer corpus, we built an open vocabulary background collection model to deal with out of vocabulary words by smoothing the unigram probabilities with Good-Turing discounting, using the SRILM toolkit[8] (Stolcke, 2002).

5.3 Results

As baselines, we consider the query-likelihood model (QLM), corresponding to equation 4 with $\beta = 0$, and Lucene.[9]

The results reported in Table 4 show that models incorporating monolingual translation probabilities perform consistently better than both baseline systems especially when they are used in combination. It is however difficult to provide a ranking of the different types of training data based on the retrieval results: it seems that LSR is slightly more performant than WAQ and WAQA, both alone and in combination, but the improvement is minor. It is worth noticing that while the LSR training data are comparatively smaller than WAQ and WAQA, they however yield comparable results. The linear combination of datasets (WAQ+WAQA+LSR$_{\text{Lin}}$) yields statistically significant performance improvement when compared to the models without combinations (except when compared to WAQA for R-Prec$_1$, $p>0.05$), which shows that the different datasets and resources used are complementary and each contribute to the overall result.

Three answer retrieval examples are given in Figure 1. They provide further evidence for the results obtained. The correct answer to the first question "Who invented Halloween?" is retrieved by the WAQ+WAQA+LSR$_{\text{Lin}}$ model, but not by the QLM. This is a case of a weak match with only "Halloween" as matching term. The WAQ+WAQA+LSR$_{\text{Lin}}$ model is however able to establish the connection between the question term "invented" and the answer term "originated". Questions 2 and 3 show that translation probabilities can also replace word normali-

[8] http://www.speech.sri.com/projects/srilm/

[9] http://lucene.apache.org

QLM top answer	**WAQ+WAQA+LSR**$_{\text{Lin}}$ top answer
Question 1: Who invented **Halloween**?	
Halloween occurs on October 31 and is observed in the U.S. and other countries with masquerading, bonfires, and games.	The observances connected with **Halloween** are thought to have originated among the ancient Druids, who believed that on that evening, Saman, the lord of the dead, called forth hosts of evil spirits.
Question 2: Can **mosquito** bites **spread AIDS**?	
Another species, the Asian tiger **mosquito**, has caused health experts concern since it was first detected in the United States in 1985. Probably arriving in shipments of used tire casings, this fierce biter can **spread** a type of encephalitis, dengue fever, and other diseases.	Studies have shown no evidence of HIV transmission through insects – even in areas where there are many cases of **AIDS** and large populations of insects such as *mosquitoes*.
Question 3: How do the **mountains form** into a **shape**?	
In 1985, scientists vaporized graphite to produce a stable **form** of carbon molecule consisting of 60 carbon atoms in a roughly spherical **shape**, looking like a soccer ball.	Geologists believe that most **mountains** are *formed* by movements in the earth's crust.

Figure 1: Top answer retrieved by QLM and WAQ+WAQA+LSR$_{\text{Lin}}$. Lexical overlaps between question and answer are in bold, morphological relations are in italics.

sation techniques such as stemming and lemmatisation, since the answers do not contain the question terms "mosquito" (for question 2) and "form" (for question 3), but only their inflected forms "mosquitoes" and "formed".

6 Conclusion and Future Work

We have presented three datasets for training statistical word translation models for use in answer finding: question-answer pairs, manually-tagged question reformulations and glosses for the same term extracted from several lexical semantic resources. It is the first time that the two latter types of datasets have been used for this task. We have also provided the first intrinsic evaluation of word translation probabilities with respect to human relatedness rankings for reference word pairs. This evaluation has shown that, despite the simplicity of the method, monolingual translation models are comparable to concept vector semantic relatedness measures for this task. Moreover, models based on translation probabilities yield significant improvement over baseline approaches for answer finding, especially when different types of training data are combined. The experiments bear strong evidence that several datasets encode different and complementary types of knowledge, which are all useful for retrieval. In order to integrate semantics in retrieval, it is therefore advisable to combine both knowledge specific to the task at hand, e.g. question-answer pairs, and external knowledge, as contained in lexical semantic resources.

In the future, we would like to further evaluate the models presented in this paper for different tasks, such as question paraphrase retrieval, and larger datasets. We also plan to improve question analysis by automatically identifying question topic and question focus.

Acknowledgments We thank Konstantina Garoufi, Nada Mimouni, Christof Müller and Torsten Zesch for contributions to this work. We also thank Mark-Christoph Müller and the anonymous reviewers for insightful comments. We are grateful to Bill Dolan for making us aware of the Microsoft Research QA Corpus. This work has been supported by the German Research Foundation (DFG) under the grant No. GU 798/3-1, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

References

Adam Berger and John Lafferty. 1999. Information Retrieval as Statistical Translation. In *Proceedings of the 22nd Annual International Conference on Re-*

search and Development in Information Retrieval (SIGIR '99), pages 222–229.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding. In *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR '00)*, pages 192–199.

Hui Fang. 2008. A Re-examination of Query Expansion Using Lexical Resources. In *Proceedings of ACL-08: HLT*, pages 139–147, Columbus, Ohio.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing Search in Context: the Concept Revisited. *ACM Transactions on Information Systems (TOIS)*, 20(1):116–131, January.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1606–1611.

Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural Language Based Reformulation Resource and Wide Exploitation for Question Answering. In *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding Similar Questions in Large Question and Answer Archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, pages 84–90.

Thomas K. Landauer, Darrell Laham, and Peter Foltz. 1998. Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. *Advances in Neural Information Processing Systems*, 10:45–51.

Jung-Tae Lee, Sang-Bum Kim, Young-In Song, and Hae-Chang Rim. 2008. Bridging Lexical Gaps between Queries and Questions on Large Online Q&A Collections with Compact Translation Models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 410–418, Honolulu, Hawaii.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208.

Christof Müller, Iryna Gurevych, and Max Mühlhäuser. 2007. Integrating Semantic Knowledge into Text Similarity and Information Retrieval. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC)*, pages 257–264.

Vanessa Murdock and W. Bruce Croft. 2005. A Translation Model for Sentence Retrieval. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, pages 684–691.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Yonggang Qiu and Hans-Peter Frei. 1993. Concept Based Query Expansion. In *Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR '93)*, pages 160–169.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical Machine Translation for Query Expansion in Answer Retrieval. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL' 07)*, pages 464–471.

Stefan Riezler, Yi Liu, and Alexander Vasserman. 2008. Translating Queries into Snippets for Improved Query Expansion. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 737–744.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904.

Noriko Tomuro. 2003. Interrogative Reformulation Patterns and Acquisition of Question Paraphrases. In *Proceedings of the International Workshop on Paraphrasing*, pages 33–40.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval Models for Question and Answer Archives. In *Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval (SIGIR '08)*, pages 475–482.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In *Data Structures for Linguistic Resources and Applications*, pages 197–205. Gunter Narr, Tübingen.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI 2008)*, pages 861–867.

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical Query Paraphrasing for Document Retrieval. In *Proceedings of the 19th International Conference on Computational linguistics*, pages 1177–1183, Taipei, Taiwan.