Cross-Lingual Question Answering Using Off-the-Shelf Machine Translation

Kisuh Ahn, Beatrice Alex, Johan Bos, Tiphaine Dalmas, Jochen L. Leidner, and Matthew B. Smillie

University of Edinburgh, Scotland, UK
treq-qa@inf.ed.ac.uk

Abstract. We show how to adapt an existing monolingual open-domain QA system to perform in a cross-lingual environment, using off-the-shelf machine translation software. In our experiments we use French and German as source language, and English as target language. For answering factoid questions, our system performs with an accuracy of 16% (German to English) and 20% (French to English), respectively. The loss of correctly answered questions caused by the MT component is estimated at 10% for French, and 15% for German. The accuracy of our system on correctly translated questions is 28% for German and 29% for French.

1 Introduction

In this paper we investigate the use of off-the-shelf machine translation (MT) software to adapt monolingual automatic question answering (QA) to perform in a cross-lingual situation. We will describe QED, a question answering system developed at the University of Edinburgh [1], and its performance on two cross-lingual QA tasks organised by the Cross Language Evaluation Forum (CLEF-2004).

QED was originally developed for monolingual (English) QA tasks, and our aim was to turn it into a cross-lingual system with a minimum of required changes. The obvious way to do this is by adding an MT component to the front-end of the system, with English as target language. We concentrated on the languages French and German for the cross-language QA task, resulting in a QA system that responds to German or French questions with English answers. So we only required an MT component to translate the questions.

The CLEF evaluation exercise for QA is based on that of TREC [2]. In short, the task is to give answers as exact as possible for factoid and definition questions, and back these up with a document that supports the answer. Questions for which no answer can be found in the document collection have to be answered with the string "NIL". Each answer needs to be associated with a confidence value (a number between 0 and 1), in order to reward systems that are able to model their own performance.

We have organised this paper as follows. First, we describe the general architecture of the cross-lingual QED question answering system as well as its

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 446–457, 2005.

individual components (Section 2). In Section 3, we present our results obtained in the CLEF-2004 evaluation, give a detailed error analysis of the MT component, and compare the performance of the cross-lingual with the monolingual task. We summarise our work and conclude in Section 4.

2 The QED System

2.1 Architecture

QED is a system originally designed for monolingual (English) QA tasks [1]. It has a traditional sequential QA architecture. From a bird's eye view, it consists of question analysis, document retrieval, and answer selection. Most of the QED system as used in this paper is similar to that described in our earlier work [1], minus the more elaborate question-typing, the use of Lemur instead of MG for Information Retrieval (IR), several minor enhancements in the various components, and, of course, the MT component. We used the 200 French and German questions from CLEF-2003 [3] as development data.

Figure 1 gives a detailed overview of QED's architecture. After the questions are translated from the source language (German and French) into the target language (English), they are tokenized and possibly reformulated to increase the precision of parsing. After stemming and part-of-speech (POS) tagging, the question is parsed. A semantic representation is generated from the grammatical relations, which is used to construct a query for the document retrieval module to obtain documents.

A passage segmenting and ranking tool is used to prune the search space and find document regions likely to contain answers. Its output is parsed and a semantic representation for answer candidates is created likewise. An answer extraction module attempts to match and score representations of question and answer candidates. Finally, evidence from the Web in the form of co-occurrence counts is used to check answer candidates for validity and the best answer is output.

This is QED's architecture in a nutshell. We will consider some of these components in more detail in the following sections. We will illustrate our approach to machine translation, passage selection, question typing, linguistic analysis, semantic interpretation, and finally answer selection.

2.2 Machine Translation

Our translation component is built around Babelfish[1], an online MT engine based on Systran. This is a rule-based MT engine, which makes use of both bilingual dictionaries and linguistic rules designed empirically for specific language pairs. In order to assess the quality of a pure off-the-shelf component, we ran an experiment by translating 200 CLEF-2003 questions from German to English and judge the results for acceptability. Perhaps unsurprisingly, we ini-

[1] `http://babelfish.altavista.com/`

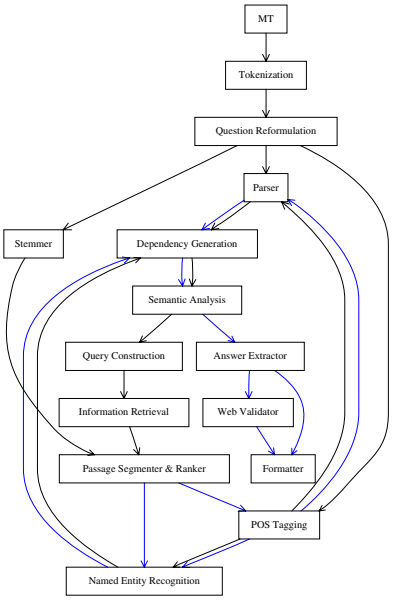

Fig. 1. The QED system architecture for CLEF-2004 (dataflow graph). Normal arrows represent processing of the question, bold arrows represents processing of answers

tially observed several translation mistakes. Only 29% of the translations were judged acceptable. Many of the errors were caused by foreign words and literally translated named entities.

However, we also noticed that the majority of the errors were systematic, and we decided to develop pre- and post-processing rules to improve the quality of the MT output. As the English MT output serves as input into the QA system, our aim was to produce MT output as correct as possible. We therefore invested some time in examining the types of errors that occurred in the Systran output for both language pairs (German to English, and French to English) and devised language-pair specific reformulation rules. We found that the best results were achieved by employing a set of meaning preserving reformulation rules before MT, as well as applying a set of rules after MT. We refer to these sets of rules as pre-MT and post-MT rules, respectively. Both sets of rules were implemented using Perl's regular expression matching techniques.

The pre-MT rules mostly reformulate certain types of questions not covered by the MT component into simpler constructions which it can actually deal with. For example, we reformulated French questions starting with *À quel moment* into questions beginning with *Quand*. In total, we created 24 pre-MT rules for French, and 9 for German.

The post-MT rules deal with systematic errors encountered in the MT output. A case in point are French questions distinguished by the inversion of subject pronoun and verb, such as *Où X travaille-t-il?*. The English MT output for this type of question is *Where X does it work?* instead of *Where does X work?*.

Another case in point are German questions such as *Wie heißt X?*, which are literally translated into *How is X called?* rather than *What is X called?*. The surface pattern-oriented pre-MT and post-MT rules enabled us to correct such errors automatically. We implemented 24 post-MT rules for French, and 25 for German.

These pre-MT and post-MT rules improved the MT component considerably, although the results were far from perfect. However, we expected them to be good enough for the cross-lingual QA challenge.

2.3 Document Retrieval, Passage Extraction and Ranking

We used the Lemur toolkit[2] to realise document retrieval based on a Vector-Space Model. The question was analysed syntactically and semantically and a weighted set of phrases was constructed from the Discourse Representation Structures (see Section 2.6), which were converted into structured queries for Lemur. The most relevant 300 documents were retrieved for subsequent processing.

Our passage segmentation and ranking component takes a query and a set of retrieved documents and extracts n-sentence passages (called "tiles"), and assigns a score to them. This is done by sliding an n-sentence window over the document stream (where we set n=3, as this gave the best results in training), retaining all window tiles that contain *at least one* of the words in the query and also always must contain *all* upper-case query words. The score is based on heuristic rules based on the following features:

- the number of non-stopword query word tokens (as opposed to types) present in the tile;
- a comparison of the capitalization of query occurrence and tile occurrence of a term;
- the occurrence of bigrams and trigrams in both question and tile.

Each tile's score is multiplied with a slightly asymmetric triangular window function to weight sentences in the centre of a window higher than in the periphery and to break ties. The output of the tiler is the top-scoring 100 tiles (eliminating duplicates). More information on this component can be found in our earlier work [1].

2.4 Question Typing

We used a hierarchical taxonomy of eleven basic question types (Fig. 2), based on the strategies used for finding suitable answers within the large variety of question patterns. This division is based on answers in the form of the linguistically motivated categories S (sentence), ADJ (adjective) and NP (noun phrase). Some of the question-types are further divided into subtypes, where C is a concept, R a relation, and U a unit of measurement. Note that although there are only

[2] `http://www-2.cs.cmu.edu/~lemur/`

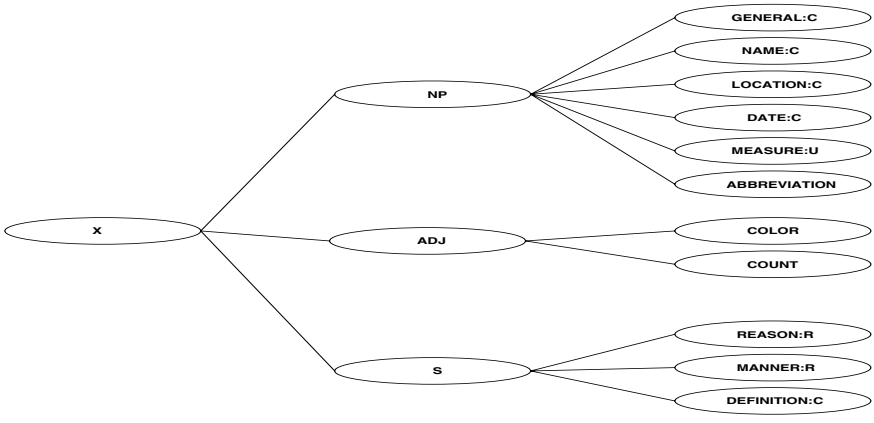

Fig. 2. The Question Type Taxonomy used in QED

eleven basic types, the values of the subtype parameters allow us to generate an infinite number of question types.

The question types are determined after the semantic analysis of the question using a rule-based system. For instance, *How hot is the sun?* is assigned the question type MEASURE:TEMPERATURE, and *Who is Janis Joplin?* the question type DEFINITION:PERSON. The question types are used by the answer selection component to constrain the set of potential answers.

2.5 Linguistic Analysis

The C&C maximum entropy POS tagger [4] is used to tag the question words and the text segments returned by the tiler. The C&C named entity tagger [5] is also applied to the question and text segments, identifying named entities from the standard MUC-7 data set (locations, organisations, persons, dates, times and monetary amounts). The POS tags and named entity tags are used to assist semantic interpretation (see Section 2.6).

We used the Radisp system [6] to parse the question and the text segments returned by the tiler. The Radisp parser returns syntactic dependencies represented by grammatical relations such as NCSUBJ (non-clausal subject), DOBJ, (direct object), NCMOD (non-clausal modifier), and so on. The set of dependencies for a sentence are annotated with POS and named entity information and converted into a graph in Prolog format.

The parser's performance on questions is not fantastic, probably because it is trained on newspaper texts. To increase the quality of the parser's output for questions, we reformulated questions in imperative form (e.g. *Name countries in Europe*) into interrogative form (*What are countries in Europe?*), and applied this reformulation technique to other question types not handled well by the parser. The Radisp parser was much better at returning the correct dependencies for these reformulated questions.

The output of the parser, a graph describing a set of dependency relations between syntactic categories, is used to build a semantic representation—both for the question under consideration and for the text passages that might contain an answer to the question. Categories contain the following information: the surface word-form, the lemmatized word-form, the word position in the sentence, the sentence position in the text, named-entity information, and a POS tag defining the category.

2.6 Semantic Interpretation

Our semantic formalism is based on Discourse Representation Theory [7], but we use an enriched form of Discourse Representation Structure (DRS), combining semantic information with syntactic and sortal information. DRSs are constructed from the dependency relations in a recursive way, starting with an empty DRS at the top node of the dependency graph, and adding semantic information to the DRS as we follow the dependency relations in the graph, using the POS information to decide on the nature of the semantic contribution of a category.

Following Discourse Representation Theory, a DRS is defined as an ordered pair of a set of discourse referents and a set of DRS-conditions. We consider the following types of DRS-conditions: `pred(x,S)`, `named(x,S)`, `card(x,S)`, `event(e,S)`, and `argN(e,x)`, `rel(x,y,S)`, `mod(x,S)`, where `e`, `x`, `y` are discourse referents, `S` a constant, and `N` a number between 1 and 3, designating an abstract semantic role. Questions introduce a special DRS-condition of the form `answer(x,T)` for a question type `T`, called the the *answer literal*. Answer literals play an important role in answer selection (see Section 2.7).

Implemented in Prolog, we reached a recall of around 80%. (By *recall* we mean the percentage of categories that contributed to semantic information in the DRS.) Note that each passage or question is translated into one single DRS, hence DRSs can span several sentences. Some basic techniques for pronoun resolution are implemented as well. However, to avoid complicating the answer extraction task too much, we only considered non-recursive DRSs in our implementation, i.e. DRSs without complex conditions introducing nested DRSs for dealing with negation, disjunction, or universal quantification.

Finally, a set of DRS normalisation rules are applied in a post-processing step, thereby dealing with active-passive alternations, question typing, inferred semantic information, and the disambiguating of noun-noun compounds. The resulting DRS is enriched with information about the original surface word-forms and POS tags, by co-indexing the words, POS tags, the discourse referents, and DRS-conditions.

2.7 Answer Selection

The answer extraction component takes as input a DRS for the question, and a set of DRSs for selected passages. The task of this component is to extract answer candidates from the passages. This is realised by performing a match between the question-DRS and a passage-DRS, by using a relaxed unification

method and a scoring mechanism indicating how well two DRSs match each other.

Taking advantage of Prolog unification, we use Prolog variables for all discourse referents in the question-DRSs, and Prolog atoms in passage-DRSs. We then attempt to unify all terms of the question DRSs with terms in a passage-DRS, using an A* search algorithm. Each potential answer is associated with a score, which we call the DRS-score. High scores are obtained for perfect matches (i.e., standard unification) between terms of the question and passage, low scores for less perfect matches (i.e., obtained by "relaxed" unification). Less perfect matches are granted for different semantic types, predicates with different argument order, or terms with symbols that are semantically familiar according to WordNet [8].

After a successful match the answer literal is identified with a particular discourse referent in the passage-DRS. Recall that the DRS-conditions and discourse referents are co-indexed with the surface word-forms of the source passage text. This information is used to generate an answer string, simply by collecting the words that belong to DRS-conditions with discourse referents denoting the answer. Finally, all answer candidates are output in an ordered list. Duplicate answers are eliminated, but answer frequency information is added to each answer in this final list.

3 Evaluation and Results

3.1 Results at the CLEF-2004 Campaign

We submitted two runs for each language pair, differing in the way reranking of answers was executed. We considered two reranking parameters: S, the normalised DRS-score, and F, the normalised frequency. The answers of the first runs for each language pair (`edin041deen` and `edin041fren`) were ranked using the formula $Rank = 0.2{*}S{+}0.8{*}F$, the answers of the second runs (`edin042deen` and `edin042fren`) were ranked using the formula $Rank = 0.8 * S + 0.2 * F$ for location and measure question types, and on $Rank = 1.0 * S$ for all other question types. The weights were estimated on the basis of running QED on TREC-2003 data.

For both languages, the second runs performed the best (as expected), with an overall accuracy of 17.00% for German and 20.00% for French. The results for the factoid and definition questions are listed in Table 1 and Table 2.

Table 1. CLEF-2004 Performance of QED on Factoid Questions

Run	Right	Inexact	Unsupported	Accuracy
`edin041deen`	24	4	1	13.33%
`edin042deen`	29	5	0	16.11%
`edin041fren`	32	4	0	17.78%
`edin042fren`	37	6	0	20.56%

Table 2. CLEF-2004 Performance of QED on Definition Questions

Run	Right	Inexact	Unsupported	Accuracy
edin041deen	4	1	0	20.00%
edin042deen	5	2	0	25.00%
edin041fren	1	2	0	5.78%
edin042fren	3	1	0	15.00%

For the German edin041deen and edin042deen runs, the answer-string "NIL" was returned 47 times, and correctly returned 7 times (14.89%). For the French edin041fren and edin042fren, the answer-string "NIL" was returned 70 times, and correctly returned 11 times (15.71%). The confidence-weighted score for the four runs varied between 0.04922 and 0.05889, which is low compared to other systems, and indicates that tbere is a lot of room for improvement on self-assessment in QED.

3.2 Measuring Impact of MT

After the CLEF-2004 campaign we ran several more experiments to assess the impact of the errors introduced by the MT component. Both the French and German questions were translated from the same set of source English questions. Running the QED system on these English questions, surpassing the MT component, would give us concrete information in terms of performance loss when using off-the-shelf MT in cross-lingual QA.

Obviously, there are some problems with evaluating the results compared to the evaluation at the CLEF campaign. It is difficult to get objective judgements for exactness, and to a certain extent this also holds for the documents that support the answers. To overcome these difficulties, we compared the results of answers comprising all correct, inexact, and unsupported answers. Also, we didn't consider NIL answers in the comparison, because the relatively high number of correct NIL answers for the French run would bias the comparison considerably. We used the list of all correct answers generated by all entries of CLEF-2004 for our judgements.

The results of this experiment were interesting. For the English to English configuration, the total of correctly answered questions was 40. For French to English, the number of correct answers was 36, indicating a loss of only 10%. For German to English, the number of correct answers was 34, corresponding to a drop of 15%. Therefore, the loss of answers introduced by the MT component was reasonably low.

3.3 Error Analysis of Question Translation

In order to gain a better understanding as to where MT errors occur and how to improve the system, we performed an error analysis of the translated CLEF-2004 questions. The types of errors in the output of the MT component can be classed into nine separate categories. We will present these categories and give

examples of each (some of them are hilarious, but they illustrate the difficulties in MT).

1. **Content Word**
   DE: Nenne einen Grund für Selbstmord bei Teenagern.
   EN: Name a reason for suicide with dte rodents.
2. **Word Order**
   FR: En quelle année les jeux Olympiques ont eu lieu à Barcelone?
   EN: In which year the Olympic Games did take place in Barcelona?
3. **Untranslated Word**
   FR: Quel animal roucoule?
   EN: Which animal roucoule?
4. **Translated Named Entity**
   DE: Was verkauft Faust dem Teufel?
   EN: What sells fist to the devil?
5. **Untranslated Named Entity**
   DE: Wo ist die Eremitage?
   EN: Where is the Eremitage?
6. **Mistranslated Named Entity**
   FR: Qui a écrit le Petit Prince?
   EN: Who wrote the Small Prince?
7. **Verb Form, Tense or Number**
   DE: Wer sind die Simpsons?
   EN: Who is the Simpsons?
8. **Missing Verb**
   FR: Qu'est-ce que l'UEFA?
   EN: What the UEFA?
9. **Minor**
   DE: Nenne eine Ölgesellschaft.
   EN: Name a oil company.

We classified all incorrectly translated question into one of these nine categories. In some cases more than one type of error occurred, in which case we picked the category which made the translation most incomprehensible. Table 3 lists the types of errors and their frequency in the English MT output that was obtained from the original 200 German and French questions. The table shows that the types of errors that occur are relatively language-specific, since the distribution of errors is very different for the two language pairs.

The main source of error for both systems (DE→EN: 27%; FR→EN: 35.5%) are wrong and awkwardly phrased translations of content words. For instance, in the above example, the noun "Teenagern" was mistakenly treated as the German compound "Tee+nagern". Moreover, the output quality of the French to English system also suffers from wrong word order for 11.5% of the questions which only happened 6.5% of the time when translating from German to English. The German to English system, however, produces considerably more errors when dealing with unknown words and named entities that should not be translated (see

Table 3. Source of MT errors and their frequency distributed over different categories, plus the number of correctly answered questions in each category

Type of Error	**DE → EN**		**Correct**	**FR → EN**		**Correct**
Content Word	54	27.0%	7	71	35.5%	13
Word Order	13	6.5%	1	23	11.5%	3
Untranslated Word	11	5.5%	0	7	3.5%	0
Translated Named Entity	8	4.0%	0	1	0.5%	0
Untranslated Named Entity	5	2.5%	0	4	2.0%	0
Mistranslated Named Entity	4	2.0%	0	5	2.5%	0
Verb Form, Tense or Number	8	4.0%	1	5	2.5%	0
Missing verb	0	0.0%	0	8	4.0%	0
Minor errors	22	11.0%	4	17	8.5%	3
Total incorrectly translated	125	62.5%	13	141	70.5%	19
Total correctly translated	75	37.5%	21	59	29.5%	17
Total	200	100.0%	34	200	100.0%	36

Table 3). The German to English system also makes more mistakes in choosing the correct verb form, tense and number. The French to English system on the other hand never translates the verb in questions beginning with "Qu-est-ce que" (What is). This is an error specific to the French-English language pair that never occurs for other language pair scenarios. The category "Minor errors" contains correct translations but with missing or wrong articles or wrong case which will not necessarily affect the performance of the QA system. Overall, the German to English MT system produces 8% more correct output than the French to English system.

Table 3 also lists the number of incorrectly translated questions for which our QA system nevertheless produced correct answers. Here, we refer to correct, inexact and unsupported answers as in the previous section. For German, 38.2% of correctly answered questions (13 out of 34) contain translation mistakes, including 4 questions with minor errors. For French, this percentage is considerably higher at 52.8% (19 out of 36) and includes 3 questions with minor errors.

Even though the output of the French to English MT system is of significant lower quality, it yields better QA scores than in the German to English scenario. One of the reasons for this seeming inconsistency is the fact that translation errors vary in severity. It appears that QED is still able to produce correct answers for some questions with incorrectly or awkwardly translated content words. Despite these errors, such questions still provide sufficient information and are therefore easier to answer than questions with wrong named entities, an error which was made more frequently by the German to English MT system.

Interestingly, the ratio of correctly answered to correctly translated questions is approximately the same for both languages (28.0% for German, and 28.8% for French). However, the ratio of correctly answered to incorrectly translated questions is only around 10% for the German to English system and 13% for the French to English system. This clearly shows that by further improving

the quality of the MT output, the performance of the QA system can still be increased.

For future work, we suggest using several competing MT systems in a parallel architecture. Automatic MT evaluation scores like Bleu [9] could also be considered to select the best translation from a set of candidate translations if multiple engines are available. Questions translated by multiple MT systems could be used together as query expansions. Another proposed extension is recognition (and alignment) of Named Entities in source and target questions to avoid literal translations of proper nouns (for instance, *Spielberg→play mountain* and *Neufeld→new field*).

4 Conclusion

We have presented extensions to a mono-lingual QA system to enable it for a cross-lingual task. Our approach consisted of composing existing software (with minor enhancements) for machine translation and question answering in a sequential pipeline. The translation was enhanced using pattern replacements to correct systematic mistakes. We obtained an accuracy of 16% (German to English) and 20% (French to English), respectively, for answering factoid questions. For definition questions, we obtained an accuracy of 25% (German to English) and 15% (French to English), respectively. Definition questions constituted a minor portion of the test set.

We showed that it is feasible to use out-of-the-box machine translation software to transform a monolingual QA system into a multilingual one. Despite the large number of translation mistakes, the majority do not affect the overall result of question answering, and some simple pre- and postprocessing rules can successfully deal with systematic errors. For the questions at the CLEF-2004 campaign, the loss of correct answers for French to English was only 10%, and for German to English 15%, compared to English to English processing. Only considering correctly translated questions, the accuracy of the system was 28% for German and 29% for French on factoid and definition questions.

Acknowledgements

We are grateful to Steve Clark, James Curran, Malvina Nissim, and Bonnie Webber for assistance and helpful discussions, and would like to thank the system administrators Bill Hewitt and Andrew Woods for their computing support. Special thanks go to John Carroll for his help with the Radisp parser, and in general to all developers of all external programs that we used in QED. We also would like to thank Bernardo Magnini, Carol Peters, Maarten de Rijke (in particular for supplying us with a crucial password at a crucial time), and Alessandro Vallin for all organisational CLEF issues they dealt with so adequately.

Alex is supported by Scottish Enterprise Edinburgh-Stanford Link (R36759), the Economic and Social Research Council, UK and the School of Informatics, University of Edinburgh. Dalmas is supported by the School of Informatics, Uni-

versity of Edinburgh. Leidner is supported by the German Academic Exchange Service (DAAD) under scholarship D/02/01831 and by Linguit GmbH (research contract UK-2002/2).

References

1. Leidner, J.L., Bos, J., Dalmas, T., Curran, J.R., Clark, S., Bannard, C.J., Steedman, M., Webber, B.: The QED open-domain answer retrieval system for TREC 2003. In: Proceedings of the Twelfth Text Retrieval Conference (TREC 2003). NIST Special Publication 500-255, Gaithersburg, MD (2004) 595–599.
2. Voorhees, E.M.: Overview of TREC 2003. In: Proceedings of the Twelfth Text Retrieval Conference (TREC 2003). NIST Special Publication 500-255, Gaithersburg, MD (2004) 1–13.
3. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., de Rijke, M.: Creating the DISEQuA corpus: a test set for multilingual question answering. In Peters, C., ed.: Working Notes for the CLEF 2003 Workshop, Trondheim, Norway (2003).
4. Curran, J.R., Clark, S.: Investigating GIS and smoothing for maximum entropy taggers. In: Proceedings of the 11th Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest, Hungary (2003) 91–98.
5. Curran, J.R., Clark, S.: Language independent NER using a maximum entropy tagger. In: Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03), Edmonton, Canada (2003) 164–167.
6. Briscoe, T., Carroll, J.: Robust accurate statistical annotation of general text. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria (2002) 1499–1504.
7. Kamp, H., Reyle, U.: From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT. Kluwer, Dordrecht (1993).
8. Fellbaum, C., ed.: WordNet. An Electronic Lexical Database. The MIT Press (1998).
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Thomas J. Watson Research Center (2001).