Monolingual and Multilingual Question Answering on European Legislation

Radu ION, Alexandru CEAUŞU, Dan ŞTEFĂNESCU, Dan TUFIŞ, Elena IRIMIA and Verginica BARBU MITITELU

Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie no. 13, Bucharest 050711, Romania
{radu, aceausu, danstef, tufis, elena, vergi}@racai.ro

This paper documents the participation of the Research Institute for Artificial Intelligence to the CLEF 2010 ResPubliQA lab. We answered questions in Romanian and English from Romanian documents of Acquis Communautaire and the European Parliament Proceedings. We extend the report from the previous ResPubliQA participation by introducing multi-factored paragraph relevance score training onto English-Romanian QA. We also investigate how our monolingual parametric QA system developed for the last year's ResPubliQA track scales up to current challenges.

1. Introduction

Research Institute for Artificial Intelligence (RACAI) is at the 5th participation in the CLEF (http://www.clef-campaign.org/) series of Question Answering systems evaluation. We have built Question Answering (QA) systems for the English-Romanian or the Romanian-Romanian tracks experimenting with each passing year. Beginning with last year, the QA task simplified in that the organizers asked for the single most relevant paragraph containing the answer to the user's natural language question. Also, questions were independent one from the other and no anaphora resolution was required in order to find referents of the question pronouns in previous questions and/or answers. Thus, the road for a reliable QA system development was opened and continues to be opened for the 2010 edition of the popular QA systems evaluation forum.

This year we participated to the Romanian-Romanian track of the ResPubliQA lab as we did in 2009, but we also enrolled in the English-Romanian cross-lingual QA in order to check some hypotheses that we put forward [2]. The approach that we took was to use the last year's test set comprised of 500 questions from the JRC Acquis corpus in order to train our paragraph relevance weights. The difference is that for ResPubliQA 2010, the Europarl corpus was added along with a new type of question for which we did not have any training data. This type of question (dubbed OPINION in the "ResPubliQA 2010 - Track Guidelines" document) was specific to the Europarl corpus in which each speaker in the European Parliament expresses an opinion about a given state of affairs.

In what follows, we will describe the document collection, our QA systems (both monolingual and cross-lingual) and the results we have obtained. In doing so, we will be brief on subjects that have already been presented at length elsewhere [2], focusing on new aspects and discussions pertaining to the task at hand.

2. The Document Collection

The document collection was composed from two corpora: the JRC Acquis [9] that was introduced last year and the new addition of Europarl [4]. The latter corpus consists of 142 documents in both English and Romanian containing almost 8.6M tokens in English and almost 9.3M tokens in Romanian (including punctuation). Both parts of the corpus were preprocessed using the TTL web service [12] to obtain POS tagging, lemma and chunking information (the same annotations as for the JRC Acquis corpus). As with the JRC Acquis corpus, we paragraph-aligned the Europarl corpus using the 1:1 sentence aligner developed by Moore [6]. We managed to obtain a percent of 98.76% English paragraphs that were 1:1 aligned to Romanian paragraphs which means that the corpus was already "almost" aligned with paragraph counts differing very little between English or Romanian parts for each pair of documents.

For this year's ResPubliQA competition, the JRC Acquis and the Europarl corpora were also word sense disambiguated using one of the algorithms with which we participated to the "Task #17: All-words Word Sense Disambiguation (WSD) on a Specific Domain" of the SemEval-2010 semantic evaluations forum[1]. We wanted to evaluate the impact of WSD onto the accuracy of our QA system by doing an informed, WSD-driven query expansion and WSD-enhanced document retrieval. The algorithm that we used to sense-annotate the document collection was the variant of RACAI-1 [1] which outputs the best two senses for the target word and whose reported accuracy is around 82.5% if it is applied onto highly domain-specific content words.

Figure 1 exemplifies the level of corpus annotation used by our present QA system:

```
<w lemma="the" ana="Dd" chunk="Np#21">the</w>
<w lemma="proposal" ana="Ncns" chunk="Np#21" wns="ili:ENG20-06719629-n,ENG20-06720394-n">proposal</w>
<w lemma="for" ana="Sp" chunk="Pp#9">for</w>
<w lemma="a" ana="Ti-s" chunk="Pp#9,Np#22">a</w>
<w lemma="council" ana="Ncns" chunk="Pp#9,Np#22" wns="ili:ENG20-07808337-n,ENG20-07809840-n">Council</w>
<w lemma="decision" ana="Ncns" chunk="Pp#9,Np#22" wns="ili:ENG20-05500743-n,ENG20-04628484-n">decision</w>
```

Fig. 1. The annotations in a ready to be indexed file of the document collection

[1] http://semeval2.fbk.eu/semeval2.php

In order to check for the query expansion benefits, for each sense disambiguated word, we also indexed its synonyms as given by respective ILIs[2]. For instance, for the lemma "proposal", the index contained the synonym "proposition" since this literal appears next to "proposal" in the synset which is identified by ILI "ENG20-06719629-n".

3. The QA System

The QA system has no significant modifications since the ResPubliQA 2009 [2]. It is based on a flow of web services that takes a user's natural language question, preprocess it on the fly to obtain all the annotations from Figure 1, transforms it into a Boolean query using one of the two query generation algorithms [2; pages 7, 8] and then retrieves a list of relevant paragraphs that are very likely to contain the answer to the user's question.

The way in which the paragraph list is sorted (in order to extract and return the first paragraph as the single answer to the question) is the key to the *trainable* quality of our QA system. Thus, the sort key S of a paragraph p is a linear combination of paragraph relevance scores:

$$S(p) = \sum_i \lambda_i s_i, \quad \sum_i \lambda_i = 1 \tag{1}$$

where the weights sum to 1 and are estimated by the following MERT-like procedure [7]: given a training set of questions for which the correct paragraphs are known, run the QA system for all possible values of weights such that the increment step is 0.01 and compute the MRR of each run. Retain that set of weights for which the MRR is the highest.

In order to comply with the organizers' suggestion that an "I don't know/I'm not sure" answer (identified with NOA – standing for "NO Answer") is better than a wrong answer, we introduced the combined QA system which considers the outputs of the two different query generation algorithms in the following manner:

$$\arg\min_p \left(rank_1(p) + rank_2(p)\right), \quad rank_1(p) \le K, rank_2(p) \le K, K \le 50 \tag{2}$$

where $rank(p)$ is the rank of the paragraph p in the list returned by the search engine and subscripts 1 and 2 indicate the paragraph lists returned by the respective query generation algorithm (1 for the TF-IDF one and 2 for the chunk-based one). Intuitively, the paragraph that is preferred by the combined QA system is the lowest numbered one that is common to both lists. When no such paragraph exists, the QA system returns NOA.

[2] The Inter-Lingual Index (ILI) was a major outcome of the EuroWordNet project [13] and it ensures the cross-linguistic alignment of wordnets.

4. The Monolingual Runs

We participated in the Romanian-Romanian section of the ResPubliQA 2010 Paragraph Selection (PS) track. As in the previous year, the requirement was to return exactly one paragraph containing the correct answer to each natural language question in the 200 questions test set. If the system is not sure, the NOA answer may be returned with an option to record an actual answer (paragraph) with the NOA so that the organizers may compute additional performance figures such as the percent of correct/incorrect answers out of the NOA ones.

We have submitted two Romanian-Romanian runs: **icia101PSroro** and **icia102PSroro** with the following characteristics:

- The first run is simply the QA system from ResPubliQA 2009 with global weights training as described in that paper [2] and using the TF-IDF query generation algorithm. We wanted to see how our 2009 QA system *scales up* to current challenges *without any modifications*;
- The second run is a more elaborate one. We trained the weights of the QA system on ResPubliQA 2009 500 questions test set in order to derive a set of weights for each question class for each query generation algorithm (see tables 1 and 2). Then, we combined the outputs of the TF-IDF and CHUNK QA systems, using K=1 from eq. 2.

TF-IDF Q. Gen.	Lexical Chains	Class	BLEU	Paragraph	Document
DEFINITION	0.03	0	0	0.08	0.89
FACTOID	0.06	0	0.12	0.28	0.54
PROCEDURE	0.09	0	0.09	0.17	0.65
REASON-PURPOSE	0.48	0	0.21	0.26	0.05

Table 1. Weights for each paragraph relevance score and each question class trained from the last year's test set using the TF-IDF query generation algorithm

CHUNK Q. Gen.	Lexical Chains	Class	BLEU	Paragraph	Document
DEFINITION	0.11	0	0.12	0.1	0.76
FACTOID	0.11	0	0.27	0.14	0.48
PROCEDURE	0.59	0	0.29	0.03	0.09
REASON-PURPOSE	0	0	0.09	0.17	0.74

Table 2. Weights for each paragraph relevance score and each question class trained from the last year's test set using the CHUNK query generation algorithm

The official results of our two runs are given in Table 3.

	icia101PSroro	icia102PSroro
ANSWERED	200	92
UNANSWERED	0	108
ANSWERED with RIGHT	93	63
ANSWERED with WRONG	107	29
UNANSWERED with RIGHT	0	0
UNANSWERED with WRONG	0	0
UNANSWERED with EMPTY	0	108
Overall accuracy	0.47	0.32
c@1 measure	0.47	0.49

Table 3. Official results on Romanian-Romanian ResPubliQA 2010

Table 3 reveals the fact that MERT training procedure is rather sensitive to the training data: a c@1 measure of 0.49 is significantly lower than the one of 0.68 we obtained last year. Still, there is also the issue of the size of the test data which was 200 questions vs. 500 questions last year (more than double). This translates in a reduced margin of error.

For this year's ResPubliQA competition we also wanted to test the influence the WSD has on both document/paragraph retrieval and query generation. We have already explained how we index a term using its assigned senses.

For the query side, we opted to implement a *query expansion mechanism* based on performing WSD to the user's question and generate all synonyms from the Romanian WordNet for each semantically disambiguated term. In order to do that, we used the RACAI-1 WSD algorithm [1] with which we have obtained an 82.5% accuracy on a domain-limited lexical sample if we allowed it to output the first two senses for each target word. The query expansion algorithm works in the following way:

1. obtain a query from the natural language question using the TF-IDF query generation algorithm [2];
2. for each term in that query, apply RACAI-1 WSD algorithm (which uses a WSD model derived from the document collection and always outputs the domain-computed most frequent sense of the term according to the model – the "One Sense per Domain" hypothesis) to obtain the most probable 2 senses of the term; using Romanian WordNet, generate all its synonyms for each disambiguated sense.

Evaluating the results of the WSD-enhanced QA system, we differentiated between several types of runs: with/without query expansion and with/without WSD-enhanced indexing. We tested both query generation algorithms (TF-IDF and chunk-based) but we could only expand queries produced by the TF-IDF algorithm. Table 4 summarizes the results that we have obtained on the ResPubliQA 2010 official 200 questions test set. The QA system ran with the same global weights (see eq. 1) as per last year ResPubliQA.

Q. Gen. Algorithms	Query Expansion		No Query Expansion	
	WSD Idx.	No WSD Idx.	WSD Idx.	No WSD Idx.
TF-IDF	0.2577/ 0.3084/ 0.4690	0.2602/ 0.3054/ 0.4438	0.2914/ 0.3406/ 0.6783	**0.4020**/ **0.4748**/ 0.6934
CHUNK	–	–	0.3467/ 0.4226/ **0.7286**	0.3618/ 0.4311/ **0.7286**

Table 4. Runs to evaluate the WSD impact on our QA performance for the Romanian-Romanian setting.

Every cell in this table contains three figures separated by '/': the MRR-1 (percentage of the correct paragraphs returned only on the 1st place), the MRR (classical Mean Reciprocal Rank) and the COVERAGE (the percentage of the correct paragraphs found in the *K*=50 paragraphs list the search engine returns for each question). It clearly follows that WSD negatively impacts the performance of both IR and QA (a result that was put forward a while ago [8]). There are several explanations for this state of affairs:

- domain limited WSD works well only on a highly domain-relevant lexical sample (in mathematical terms, words that are indicative of a domain with high scores) but we sense-annotated almost all content words from the document collection in which case the WSD accuracy drops. We imposed a cut-off of 5 as the score for a domain relevant term – by comparison, the most informative term from the JRC Acquis, "emolient", has a score of 75.88;
- we extracted terms along with document relevance scores and not domain relevance scores using the TF-IDF measure and not the one we developed specially for domain-relevant terms from [1];
- question formulation follows closely the phrasing in the document collection in which case, query expansion only produces noise.

This judgment is not final since we did not experiment with any of the following:

- increase the term threshold (from 5 to what value?) when doing WSD on both the question and the document collection in order to improve WSD accuracy;
- instead of indexing synonyms for every term, we could index the term's senses ids; when doing query generation, we could expand sense ids instead of synonyms with an expected effect of noise reduction;
- train the system on the ResPubliQA 2009 500 questions test set.

5. The Cross-lingual Runs

The cross-lingual system uses the same index as the monolingual Romanian QA but for the query generation it uses an already available statistical machine translation

(SMT) system experimented in several RACAI projects [3]. The SMT system is based on Moses [5], an open source framework for rapid prototyping of machine translation systems.

The training data for the translation system consisted of the JRC Acquis corpus and EMEA - European Medicines Agency documents [10]. The abundant foreign languages fragments were filtered-out as well as the translation units with significant length differences. After filtering, the remaining corpus had a total of 1.4 million translation units. Also, the translation pairs from the Romanian Wordnet [11] aligned to the Princeton Wordnet (http://wordnet.princeton.edu/) were added to the training data.

From the training data only the lemmas of the content words were kept. The first two letters of the morpho-syntactical description were added to the lemma in order to syntactically disambiguate the terms. For example, the sentence "The medicine can only be obtained with a prescription." is transformed into the sequence: "medicine^Nc obtain^Vm prescription^Nc". Using the Moses scripts and Giza++ alignments we extracted a phrase-table (3-gram maximum length) for content-words lemmas.

We used different query translation algorithms for the two submitted runs. For the first run, we selected from the translation table the translation equivalents for each content word lemma. For example, the question "Why was Perwiz Kambakhsh sentenced to death?" is translated into the query:

de_ce Perwiz Kambakhsh (condamna OR *condamnat* OR *condamnare) (deces* OR *moarte)*

For the second run, we used the Moses decoder to generate the *n*-best translation list. The terms from the lists were collected into a single query. For example, the same English question as above is translated into the query:

de_ce Perwiz Kambakhsh condamna condamnat condamnare deces moarte fi

The two approaches have similar results as can be seen in the official results shown in Table 5:

	icia101PSenro	**icia102PSenro**
ANSWERED	197	193
UNANSWERED	3	7
ANSWERED with RIGHT candidate	58	56
ANSWERED with WRONG candidate	139	137
UNANSWERED with RIGHT candidate	0	0
UNANSWERED with WRONG candidate	0	0
UNANSWERED with EMPTY candidate	3	7
Overall accuracy	0.29	0.28
c@1 measure	0.29	0.29

Table 5. Official results for English-Romanian ResPubliQA 2010

The performance measures are significantly lower than those of the monolingual counterpart suggesting the fact that either the translation can be improved or the issue of noise introduction with alternate translations for a term is to be addressed.

6. Conclusions

RACAI participated in the Romanian-Romanian and English-Romanian settings of the ResPubliQA 2010 Paragraph Selection track using the QA system that it has developed for the last year's ResPubliQA. The main aim of our participation this year was to test the scalability of our QA system to new challenges given the fact that it performed the best last year in the Romanian-Romanian setting out of 4 runs belonging to 2 participating groups but also overall out of the 28 runs in all languages. Even if the results were lower than those of the last year, we acquired important insights on how to scale this QA system to new challenges. Thus, for instance, we validated the per-class training that gives the best results and also, we know now that MERT estimation is very sensitive to the training data set.

Acknowledgements

RACAI participation to CLEF series of QA systems evaluation was possible due to the SIR-RESDEC national project (no. D1.1.0.0.0.7/18.09.2007) that aims at developing an open-domain QA system with application to the European Legislation. This project is approaching the finish line and as such, we will leverage all our CLEF experience in order to put forward this QA system and make it available on the web.

References

1. Ion, R., and Ştefănescu, D. (2010). *RACAI: Unsupervised WSD Experiments @ SemEval-2, Task #17*. In Proceedings of the 5th International Workshop on Semantic Evaluations, SemEval-2010, pages 411—416, Uppsala, Sweden, July 15–16 2010. ACL 2010.
2. Ion, R., Ştefănescu, D., Ceauşu, Al., Tufiş, D., Irimia, E., and Barbu Mititelu, V. (2009). *A Trainable Multi-factored QA System*. In Carol Peters et al., editor, Working Notes for the CLEF 2009 Workshop, pages 14, Corfu, Greece, September, 30th - October, 2nd 2009.
3. Irimia, E., and Ceauşu, Al. (2010). *Dependency-based translation equivalents for factored machine translation*, In Alexander Gelbukh (ed.) Research In Computer Science, Special Issue on NLP and its Applications 46, pp. 205—216, ISSN: 1870-4069.
4. Koehn, Ph. (2005). EuroParl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005. Phuket, Thailand. http://people.csail.mit.edu/koehn/publications/europarl/
5. Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, Ch., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, Ch., Zens, R., Dyer, Ch., Bojar, O., Constantin, A., Herbst, E. (2007). *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
6. Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine Translation: From Research to Real Users (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135-244.
7. Och, F. J. (2003). *Minimal Error Rate Training in Statistical Machine Translation*. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167.

8. Sanderson, M. (1997). *Word Sense Disambiguation and Information Retrieval*. Technical Report (TR-1997-7), University of Glasgow, Glasgow G12 8QQ, UK, 1997.
9. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: *A Multilingual Aligned Parallel Corpus with 20+ Languages*. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 2142-2147, Genoa, Italy, May 2006. ELRA - European Language Ressources Association.
10. Tiedemann, J. (2009). *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia.
11. Tufiş, D., Ion, R., Bozianu, L., Ceauşu, Al., and Ştefănescu, D. (2008a). *Romanian WordNet: Current State, New Applications and Prospects*. In Attila Tanacs, Dora Csendes, Veronika Vincze, Christiane Fellbaum, Piek Vossen: Proceedings of 4th Global WordNet Conference, GWC-2008, University of Szeged, Hungary, January 22-25 2008, pp. 441-452.
12. Tufiş, D., Ion, R., Ceauşu, Al., and Ştefănescu, D. (2008b). *RACAI's Linguistic Web Services*. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco, May 2008. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.
13. Vossen, P. (eds) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.