Learning foci for Question Answering over Topic Maps

Alexander Mikhailian[†], **Tiphaine Dalmas**[‡] and **Rani Pinchuk**[†]
[†]Space Application Services, Leuvensesteenweg 325, B-1932 Zaventem, Belgium
{alexander.mikhailian, rani.pinchuk}@spaceapplications.com
[‡]Aethys
tiphaine.dalmas@aethys.com

Abstract

This paper introduces the concepts of *asking point* and *expected answer type* as variations of the question *focus*. They are of particular importance for QA over semi-structured data, as represented by Topic Maps, OWL or custom XML formats. We describe an approach to the identification of the question *focus* from questions asked to a Question Answering system over Topic Maps by extracting the *asking point* and falling back to the *expected answer type* when necessary. We use known machine learning techniques for *expected answer type* extraction and we implement a novel approach to the *asking point* extraction. We also provide a mathematical model to predict the performance of the system.

1 Introduction

Topic Maps is an ISO standard[1] for knowledge representation and information integration. It provides the ability to store complex meta-data together with the data itself.

This work addresses *domain portable* Question Answering (QA) over Topic Maps. That is, a QA system capable of retrieving answers to a question asked against one particular topic map or topic maps collection at a time. We concentrate on an empirical approach to extract the question *focus*. The extracted focus is then anchored to a topic map construct. This way, we map the type of the answer as provided in the question to the type of the answer as available in the source data.

Our system runs over semi-structured data that encodes ontological information. The classification scheme we propose is based on one dynamic and one static layer, contrasting with previous work that uses static taxonomies (Li and Roth, 2002).

We use the term *asking point* or AP when the type of the answer is explicit, e.g. the word `operas` in the question *What operas did Puccini write?*

We use the term *expected answer type* or EAT when the type of the answer is implicit but can be deduced from the question using formal methods. The question *Who composed Tosca?* implies that the answer is a person. That is, *person* is the *expected answer type*.

We consider that AP takes precedence over the EAT. That is, if the AP (the explicit focus) has been successfully identified in the question, it is considered to be the type of the question, and the EAT (the implicit focus) is left aside.

The claim that the exploitation of AP yields better results in QA over Topic Maps has been tested with 100 questions over the Italian Opera topic map [2]. AP, EAT and the answers of the questions were manually annotated. The answers to the questions were annotated as topic map constructs (i.e. as topics or as occurrences).

An evaluation for QA over Topic Maps has been devised that has shown that choosing APs as foci leads to a much better recall and precision. A detailed description of this test is beyond the scope of this paper.

2 System Architecture

We approach both AP and EAT extraction with the same machine learning technology based on the principle of maximum entropy (Ratnaparkhi, 1998)[3].

[1] ISO/IEC 13250:2003, `http://www.isotopicmaps.org/sam/`

[2] `http://ontopia.net/omnigator/models/topicmap_complete.jsp?tm=opera.ltm`

[3] OpenNLP `http://opennlp.sf.net` was used for tokenization, POS tagging and parsing. Maxent `http://maxent.sf.net` was used as the maximum entropy engine

	What	are	Italian	operas	?
Gold	O	O	AP	AP	O

Table 1: Gold standard AP annotation

Class	Word count	%
AskingPoint	1842	9.3%
Other	17997	90.7%

Table 2: Distribution of AP classes (word level)

We annotated a corpus of 2100 questions. 1500 of those questions come from the Li & Roth corpus (Li and Roth, 2002), 500 questions were taken from the TREC-10 questions and 100 questions were asked over the Italian Opera topic map.

2.1 AP extraction

We propose a model for extracting AP that is based on word tagging. As opposed to EAT, AP is constructed on word level not on the question level. Table 1 provides an annotated example of AP.

Our annotation guidelines limit the AP to the noun phrase that is expected to be the type of the answer. As such, it is different from the notion of focus as a noun *likely to be present in the answer* (Ferret et al., 2001) or as *what the question is all about* (Moldovan et al., 1999). For instance, a question such as *Where is the Taj Mahal?* does not yield any AP. Although the main topic is the Taj Mahal, the answer is not expected to be in a parent-child relationship with the subject. Instead, the sought after type is the EAT class LOCATION. This distinction is important for QA over semi-structured data where the data itself is likely to be hierarchically organized.

Asking points were annotated in 1095 (52%) questions out of 2100. The distribution of AP classes in the annotated data is shown in the Table 2.

A study of the inter-annotator agreement between two human annotators has been performed on a set of 100 questions. The Cohen's kappa coefficient (Cohen, 1960) was at 0.781, which is lower than the same measure for the inter-annotator agreement on EAT. This is an expected result, as the AP annotation is naturally perceived as a more complex task. Nevertheless, this allows to qualify the inter-annotator agreement as good.

For each word, a number of features were used for EAT and AP extraction.

Class	Count	%
TIME	136	6.5%
NUMERIC	215	10.2%
DEFINITION	281	13.4%
LOCATION	329	15.7%
HUMAN	420	20.0%
OTHER	719	34.2%

Table 3: Distribution of EAT classes (question level)

by the classifier, including strings and POS-tags on a 4-word window. The WH-word and its complement were also used as features, as well as the parsed subject of the question and the first nominal phrase.

A simple rule-based AP extraction has also been implemented, for comparison. It operates by retrieving the WH-complement from the syntactic parse of the question and stripping the initial articles and numerals, to match the annotation guidelines for AP.

2.2 EAT extraction

EAT was supported by a taxonomy of 6 coarse classes: HUMAN, NUMERIC, TIME, LOCATION, DEFINITION and OTHER. This selection is fairly close to the MUC typology of Named Entities[4] which has been the basis of numerous feature-driven classifiers because of salient formal indices that help identify the correct class.

We purposely limited the number of EAT classes to 6 as AP extraction already provides a fine-grained, dynamic classification from the question to drive the subsequent search in the topic map.

The distribution of EAT classes in the annotated data is shown in the Table 3.

A study of the inter-annotator agreement between two human annotators has been performed on a set of 200 questions. The resulting Cohen's kappa coefficient (Cohen, 1960) of 0.8858 allows to qualify the inter-annotator agreement as very good.

We followed Li & Roth (Li and Roth, 2002) to implement the features for the EAT classifier. They included strings and POS-tags, as well as syntactic parse information (WH-words and their complements, auxiliaries, subjects). Four lists for

[4] http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_1.html

Accuracy	Value	Std dev	Std err
EAT	0.824	0.020	0.006
Lenient AP	0.963	0.020	0.004
Exact AP	0.888	0.052	0.009
Focus (AP+EAT)	0.827	0.020	0.006

Table 4: Accuracy of the classifiers (question level)

words related to locations, people, quantities and time were derived from WordNet and encoded as semantic features.

3 Evaluation Results

The performance of the classifiers was evaluated on our corpus of 2100 questions annotated for AP and EAT. The corpus was split into 80% of training and 20% test data, and data re-sampled 10 times in order to account for variance.

Table 4 lists the figures for the accuracy of the classifiers, that is, the ratio between the correct instances and the overall number of instances. As the AP classifier operates on words while the EAT classifier operates on questions, we had to estimate the accuracy of the AP classifier per question, to allow for comparison. Two simple metrics are possible. A *lenient* metric assumes that the AP extractor performed correctly in the question if there is an overlap between the system output and the annotation on the question level. An *exact* metric assumes that the AP extractor performed correctly if there is an exact match between the system output and the annotation.

In the example *What are Italian Operas?* (Table 1), assuming the system only tagged *operas* as AP, lenient accuracy will be 1, exact accuracy will be 0, precision for the AskingPoint class will be 1 and its recall will be 0.5.

Table 5 shows EAT results by class. Tables 6 and 7 show AP results by class for the machine learning and the rule-based classifier.

As shown in Figure 1, when AP classification is available it is used. During the evaluation, AP was found in 49.4% of questions.

A mathematical model has been devised to predict the accuracy of the focus extractor on an annotated corpus.

It is expected that the focus accuracy, that is, the accuracy of the focus extraction system, is dependent on the performance of the AP and the EAT classifiers. Given N the total number of questions,

Class	Precision	Recall	F-Score
DEFINITION	0.887	0.800	0.841
LOCATION	0.834	0.812	0.821
HUMAN	0.902	0.753	0.820
TIME	0.880	0.802	0.838
NUMERIC	0.943	0.782	0.854
OTHER	0.746	0.893	0.812

Table 5: EAT performance by class (question level)

Class	Precision	Recall	F-Score
AskingPoint	0.854	0.734	0.789
Other	0.973	0.987	0.980

Table 6: AP performance by class (word level)

Class	Precision	Recall	F-Score
AskingPoint	0.608	0.479	0.536
Other	0.948	0.968	0.958

Table 7: Rule-based AP performance by class (word level)

we define the branching factor, that is, the percentage of questions for which AP is provided by the system, as follows:

$$Y = \frac{(TP_{AP} + FP_{AP})}{N}$$

Figure 1 shows that the sum AP true positives and EAT correct classifications represents the overall number of questions that were classified correctly. This accuracy can be further developed to present the dependencies as follows:

$$A_{FOCUS} = P_{AP}Y + A_{EAT}(1 - Y)$$

That is, the overall accuracy is dependent on the precision of the AskingPoint class of the AP classifier, the accuracy of EAT and the branching factor. The branching factor itself can be predicted using the performance of the AP classifier and the ratio between the number of questions annotated with AP and the total number of questions.

$$Y = \frac{(\frac{TP_{AP}+FN_{AP}}{N})R_{AP}}{P_{AP}}$$

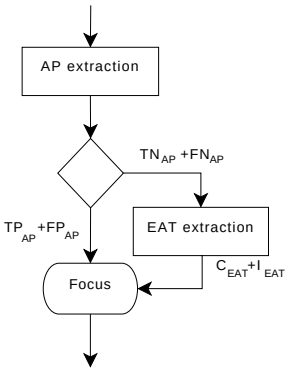

Figure 1: Focus extraction flow diagram

4 Related work

(Atzeni et al., 2004; Paggio et al., 2004) describe MOSES, a multilingual QA system delivering answers from Topic Maps. MOSES extracts a focus constraint (defined after (Rooth, 1992)) as part of the question analysis, which is evaluated to an accuracy of 76% for the 85 Danish questions and 70% for the 83 Italian questions. The focus is an ontological type dependent from the topic map, and its extraction is based on hand-crafted rules. In our case, focus extraction – though defined with topic map retrieval in mind – stays clear of ontological dependencies so that the same question analysis module can be applied to any topic map.

In open domain QA, machine learning approaches have proved successful since Li & Roth (Li and Roth, 2006). Despite using similar features, the F-Score (0.824) for our EAT classes is slightly lower than reported by Li & Roth (Li and Roth, 2006) for coarse classes. We may speculate that the difference is primarily due to our limited training set size (1,680 questions versus 21,500 questions for Li & Roth). On the other hand, we are not aware of any work attempting to extract AP on word level using machine learning in order to provide dynamic classes to a question classification module.

5 Future work and conclusion

We presented a question classification system based on our definition of *focus* geared towards QA over semi-structured data where there is a parent-child relationship between answers and their types. The specificity of the focus degrades gracefully in the approach described above. That is, we attempt the extraction of the AP when possible and fall back on the EAT extraction otherwise.

We identify the focus dynamically, instead of relying on a static taxonomy of question types, and we do so using machine learning techniques throughout the application stack.

A mathematical model has been devised to predict the performance of the focus extractor.

We are currently working on the exploitation of the results provided by the focus extractor in the subsequent modules of the QA over Topic Maps, namely anchoring, navigation in the topic map, graph algorithms and reasoning.

Acknowledgements

This work has been partly funded by the Flemish government (through IWT) as part of the ITEA2 project LINDO (ITEA2-06011).

References

P. Atzeni, R. Basili, D. H. Hansen, P. Missier, P. Paggio, M. T. Pazienza, and F. M. Zanzotto. 2004. Ontology-Based Question Answering in a Federation of University Sites: The MOSES Case Study. In *NLDB*, pages 413–420.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, No.1:37–46.

O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, and A. Vilnat. 2001. Finding an Answer Based on the Recognition of the Question Focus. In *10th Text Retrieval Conference*.

X. Li and D. Roth. 2002. Learning Question Classifiers. In *19th International Conference on Computational Linguistics (COLING)*, pages 556–562.

X. Li and D. Roth. 2006. Learning Question Classifiers: The Role of Semantic Information. *Journal of Natural Language Engineering*, 12(3):229–250.

D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus. 1999. LASSO: A Tool for Surfing the Answer Net. In *8th Text Retrieval Conference*.

P. Paggio, D. H. Hansen, R. Basili, M. T. Pazienza, and F. M. Zanzotto. 2004. Ontology-based question analysis in a multilingual environment: the MOSES case study. In *OntoLex (LREC)*.

A. Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

M. Rooth. 1992. A Theory of Focus Interpretation. *Natural Language Semantics*, 1(1):75–116.