Towards a General Model of Answer Typing: Question Focus Identification

Razvan Bunescu and Yunfeng Huang

School of EECS
Ohio University
Athens, OH 45701
bunescu@ohio.edu, yh324906@ohio.edu

Abstract. We analyze the utility of *question focus* identification for answer typing models in question answering, and propose a comprehensive definition of question focus based on a relation of coreference with the answer. Equipped with the new definition, we annotate a dataset of 2000 questions with focus information, and design two initial approaches to question focus identification: one that uses expert rules, and one that is trained on the annotated dataset. Empirical evaluation of the two approaches shows that focus identification using the new definition can be done with high accuracy, holding the promise of more accurate answer typing models.

1 Introduction and Motivation

Open domain Question Answering (QA) is one of the most complex and challenging tasks in natural language processing. While building on ideas from Information Retrieval (IR), question answering is generally seen as a more difficult task due to constraints on both the input representation (natural language questions vs. keyword-based queries) and the form of the output (focused answers vs. entire documents). A common approach to the corresponding increased complexity has been to decompose the QA task into a pipeline of quasi-independent tasks, such as question analysis, document retrieval, and answer extraction. As part of question analysis, most QA systems determine the *answer type*, i.e. *the class of the object*, or *rhetorical type of discourse*, sought by the question [1]. For example, the question Q_1*:Who discovered electricity?* is looking for the name of a HUMAN entity, whereas Q_2*:What are liver enzymes?* asks for a DEFINITION type of discourse. The corresponding answer types will therefore be HUMAN, and DEFINITION respectively. Knowledge of the answer type associated with a given question can help during the answer extraction stage, when the system can use it to filter out a wide range of candidates. Moreover, the answer type may determine the strategy used for extracting the correct answer. The HUMAN answer type for question Q_1 means that the answer is simply the name of a person, possibly identified using a named entity recognizer. A DEFINITION question like Q_2, on the other hand, may involve strategies that identify paragraphs with definition structures focused on the question topic (*liver enzymes*), or more complex strategies in which sentences on the question topic from multiple documents are automatically assembled into an answer paragraph that is given the rhetorical structure of a definition.

Most previous approaches to answer typing employ a predefined set of answer types, and use classifiers or manually crafted rules to assign answer types to questions. For example, [2] use a maximum entropy classifier to map each question into a predefined set of categories that contains all the MUC types, plus two additional categories: REASON for capturing `WHY` questions, and PHRASE as a catch-all category. Realizing the benefit of using more fine-grained categories, Li and Roth have proposed in [3] a more comprehensive set of answer types in the form of a two-level hierarchy, with a first level of 6 coarse classes that are further split into 50 fine classes on the second level. As pointed out by Pinchak and Lin in [4], using a predefined set of categories presents two major drawbacks:

1. There will always be questions whose answer types do not match any of the predefined categories, e.g. *What are the names of the tourist attractions in Reims* . Many question analysis systems employ a special catch-all category for these cases, which leads to a less effective treatment compared with the other categories.
2. The predetermined granularity of the categories leads to a trade-off between how well they match actual answer types and how easy it is to build taggers and classifiers for them. Thus, while it is relatively easy to tag names as instances of PEOPLE, this category is not a perfect fit for the question *Which former actor became president of the United States?*. Conversely, while an ACTOR answer type would be a better fit for this question, the corresponding tagging task during answer extraction will be more difficult.

As a solution to these problems, Pinchak and Lin introduced in [4] a probabilistic answer type model that directly computes the degree of fitness between a potential answer and the question context, effectively obviating the need for a predefined set of answer types. Their follow-up work in [5] presents an alternative approach to answer typing based on discriminative preference ranking. Like the probabilistic model in [4], the new flexible approach works without explicit answer types, and is shown to obtain improved results on a set of "focused" `WHAT` and `WHICH` questions.

Irrespective of whether they use an explicit set of answer types or not, many answer typing models emphasize the importance of one particular part of the question: the *question focus*, defined in [1] as "generally a compound noun phrase but sometimes a simple noun, that is the property or entity being sought by the question". According to [1], the nouns *city*, *population* and *color* are the focus nouns in the following questions:

Q_3 McCarren Airport is located in what *city*?
Q_4 What is the *population* of Japan?
Q_5 What *color* is yak milk?

The focus of a question is generally seen as determining its answer type. Singhal et al. [6], for example, use a lexicon that maps focus nouns to answer types.

They also give a syntactic rule for extracting the question focus from questions of the type *What X …*, *What is the X …*, and *Name the X …*, according to which the focus is simply the syntactic head of the noun phrase *X*. Consequently, the nouns *company* and *city* constitute the focus nouns in the following two example questions taken from [6]:

Q_6 What *company* is the largest Japanese builder?
Q_7 What is the largest *city* in Germany?

The question focus is also important in approaches that do not use explicit answer types. The models of Pinchak et al. from [4, 5] compute how appropriate an arbitrary word is for answering a question by counting how many times the word appears in *question contexts*, where a question context is defined as a dependency tree path involving the *wh*-word. For a focused question such as *What city hosted the 1988 Winter Olympics?*, the authors observed that a question focus context such as *X is a city* is more important than the non-focus context *X host Olympics*.

Motivated by the observed importance of the question focus to answer typing models, in this paper we take a closer look at the associated problem of *question focus identification*. We first give an operational definition (Section 2), followed by a set of examples illustrating the various question categories that result from a question focus analysis (Section 3). We describe a rule-based system (Section 4.2) and a machine learning approach (Section 4.3) that automatically identify focus words in input questions, and compare them empirically on a dataset of 2,000 manually annotated questions (Section 5). The paper ends with a discussion of the results and ideas for future work.

2 What is the Question Focus?

To the best of our knowledge, all previous literature on answer typing assumes that a question has at most one instance of question focus. The examples given so far in questions Q_1 through Q_7 seem to validate this assumption. Accordingly, the question focus in many questions can be extracted using simple syntactic rules such as the noun phrase (NP) immediately following the WH-word (e.g. Q_3, Q_5, Q_6), or the predicative NP (e.g. Q_4, Q_7). However, the uniqueness assumption may be violated in the case of questions matching more than one extraction rule, such as Q_6 above, or Q_8 and Q_9 below:

Q_8 Which Vietnamese *terrorist* is now a UN *delegate* in Doonesbury?
Q_9 What famed London criminal *court* was once a feudal *castle* ?

One approach to enforcing uniqueness would be to rank the extraction rules and consider only the extraction from the top most matching rule. It is unclear though on which principles the rules would be ranked. Any relative ranking between the NP immediately following the WH-word and the predicative NP seems to be arbitrary, since questions Q_8 and Q_9 can be reformulated as questions Q_{10} and Q_{11} below:

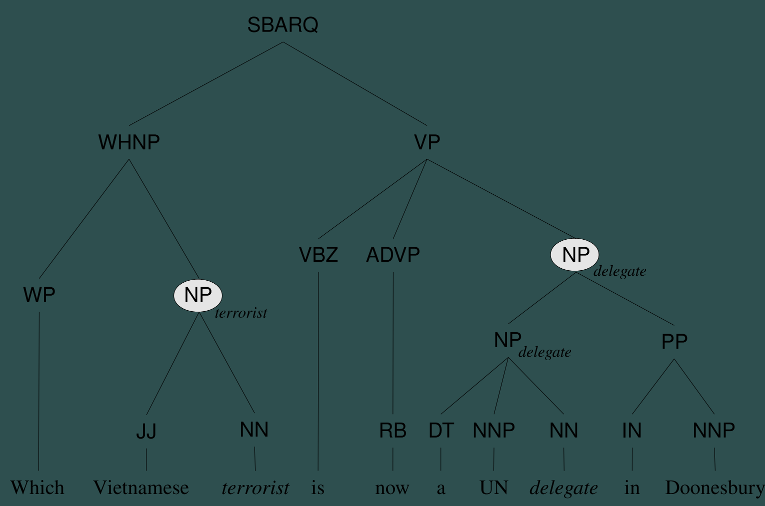

Fig. 1. Question focus example.

Q_{10} Which UN *delegate* in Doonesbury was once a Vietnamese *terrorist*?
Q_{11} What feudal *castle* is now a famed London criminal *court*?

In order to eschew these difficulties, we propose a definition of question focus that covers both the NP immediately following the WH-word and the predicative NP. For the definition to be as general as possible, it needs to take into account the fact that a question focus, as observed in [1], can denote *a property or entity* being sought by the answer. In question Q_6, for example, the focus word *company* specifies a property of the answer. In question Q_7, on the other hand, the noun phrase *the largest city in Germany* denotes the answer entity, while at the same time its head noun *city* specifies a property of the answer. Without exception, the noun phrases considered so far as potential instances of question focus have one thing in common: they can all be considered to *corefer* with the answer. It is this relation of coreference with the answer that allows us to give the following simple, yet comprehensive and operational definition for question focus:

Definition 1. *The question focus is the set of all maximal noun phrases in the question that corefer with the answer.* ■

Figure 1 shows the two noun phrases that are identified as question focus for question Q_8. The term *noun phrase* in the definition refers only to phrases marked as NP in the parse tree, and thus excludes *wh*-noun phrases (WHNP). A noun phrase is defined to be *maximal* if it is not contained in another NP with the same syntactic head.

In deriving the syntactic parse tree of a question, we use the same notation and bracketing criteria as the Penn Treebank [7], with one exception: if the WHNP contains more than a single *wh*-word, the rest of the phrase is abstracted as an NP. This contrasts with the flat structure employed in the Penn Treebank, and helps in simplifying the question focus definition.

According to the definition, a question may have one or more instances of question focus. We believe that identifying more than one focus noun phrase can be advantageous in answer extraction, when all question focus instances may be used concurrently to filter out an even wider range of answer candidates. For example, when searching for noun phrases as potential answers to question Q_8, a question answering system may choose to enforce the constraint that the noun phrase refers to both a "Vietnamese *terrorist*" and a "UN *delegate* in Doonesbury". An answer typing system that employs an explicit set of answer types such as [3] may exploit the enriched question focus (e.g. {*terrorist*, *delegate*}) to improve the accuracy of mapping questions to answer types (e.g. HUMAN). Alternatively, the identification of multiple question focus instances may also benefit approaches that do not use a predefined set of answer categories. The answer typing methods of [4, 5], for example, may choose to give preference to dependency paths that start at any focus head in the question.

There are no constraints on the type of noun phrases that may be considered as question focus. Consequently, the focus may be a definite, indefinite, or bare noun phrase, as well as a proper name, or a pronoun, as illustrated in questions Q_8 (repeated below), Q_{12}, and Q_{13}:

Q_8 Which Vietnamese *terrorist* is now a UN *delegate* in Doonesbury?
Q_{12} What French *seaport* claims to be *The Home of Wines*?
Q_{13} Who was the first black *performer* to have *his* own network TV show?

The actual semantic constraints imposed on candidate answers by a question focus vary from alternate names (e.g. *The Home of Wines*), to category information (e.g. *seaport*), to pronouns (e.g. *his*). Even a simple personal pronoun such as *his*, when identified as a question focus, may trigger a useful elimination of candidate noun phrases that do not refer to entities that are both MALE and HUMAN.

3 Question Categories

When a question has at least one instance of question focus, as question Q_{14} below, the answer type can be determined from the focus. For questions such as Q_{15} that lack an explicit question focus, the answer type is implicit in the *wh*-word if it is one of *who*, *when*, *where*, or *why*. Question Q_{16} is an example where the answer type is both implicit in the *wh*-word, and explicit in the question focus, albeit at different levels of granularity. Finally, there are questions such as Q_{17} that do not contain any explicit question focus and where the *wh*-word does not convey any information about the answer type – except maybe as a negative implicature, e.g. since the question does not use the *wh*-word *who*, then it is unlikely that the answer is of type HUMAN.

Q_{14} What *country* do the Galapagos Islands belong to?

Q_{15} Who killed Gandhi?
Q_{16} Who was the *inventor* of silly putty?
Q_{17} What do bats eat?

The implicit answer type of *how* questions is MANNER (e.g. question Q_{18}), unless the *wh*-word is followed by an adjective or an adverb, as in questions Q_{19} and Q_{20} below. A full treatment of these *quantifiable how* questions is beyond the scope of this paper (Pinchak and Bergsma [8] have recently introduced an answer typing strategy specifically designed for such cases).

Q_{18} How does a rainbow form?
Q_{19} How successful is aromatherapy?
Q_{20} How long is the Coney Island boardwalk?

Using coreference to define a question focus implies an identity relationship between the question focus and the answer, which might not be as evident for questions Q_{21}, Q_{22}, or Q_{23} below. There is nevertheless an implicit identity relationship between the focus of these questions and their answers. Taking Q_{21} as an example, the answer is a text fragment with an appropriate rhetorical structure that describes some conceptual structure X that IS the "nature of learning".

Q_{21} What is the *nature* of learning?
Q_{22} What is the *history* of skateboarding?
Q_{23} What is the *definition* of a cascade?
Q_{24} What is a cascade?

Definition questions have a special status in this category, as their answer type can be expressed either explicitly through a question focus (Q_{23}), or just implicitly (Q_{24}).

4 Automatic Identification of Question Focus

Based on the definition from Section 2, a straightforward method for solving the task of question focus identification would contain the following two steps:

1. Run coreference resolution on the question sentence.
2. Select the coreference chain that is grounded in the answer.

In this section we present a more direct approach to question focus identification, in which every word of the question is classified as either belonging to the question focus, or not, leaving the coreference resolution based approach as subject of future work.

4.1 Question Focus Dataset

In order to evaluate our word tagging approaches to question focus identification, we selected the first 2000 questions from the answer type dataset of Li and Roth [3], and for each question we manually annotated the syntactic heads of all focus instances. Since, by definition, question focus identification is a constrained version of coreference resolution, we used the annotation guidelines of the MUC 7 coreference task [9]. Three statistics of the resulting dataset are as follows:

- 1138 questions have at least one instance of question focus.
- 121 questions have two or more instances of question focus.
- 29 questions have a pronoun as one instance of the question focus.

All 29 questions that have a pronoun as a question focus also contain a non-pronominal NP focus. This property, together with the relatively low occurrence of pronouns, determined us to design the initial extraction approaches to identify only non-pronominal instances of question focus.

4.2 A Rule Based Approach

In the first approach to question focus identification, we have manually created a set of extraction rules that correspond to common patterns of focused questions. The rules, together with a set of illustrative examples, are shown in Figure 2. Given that the syntactic information is not always correct, we have decided to associate each syntactic rule with an analogous rule in which some of the syntactic constraints are approximated with part-of-speech constraints. In most cases, this meant approximating the "head of an NP" with "the last word in a maximal sequence of words tagged with one of {JJX, NNX, CD}", where JJX refers to any adjective tag, and NNX refers to any noun tag. There are in total five syntactic rules R_1 to R_5, together with their part-of-speech analogues R'_1 to R'_5. A definite noun phrase is either a noun phrase starting with a definite determiner, a possessive construction, or a proper name. Whenever the focus is extracted as the head of a possessive construction in rules R_2 and R'_2, we modify the head extraction rules from [10] to output the "possessor" instead of the "possessed" noun (e.g. output *country* as head of *country's president*). We also use two small lexicons: one for BE verbs such as {*be, become, turn into*}, and one for NAME verbs such as {*name, nickname, call, dub, consider as, know as, refer to as*}.

4.3 A Machine Learning Approach

The rules R'_i in the rule based approach were construed to complement the syntactic rules R_i by approximating noun phrase constraints with constraints on sequences of part-of-speech tags. While they often lead to the extraction of focus words that otherwise would have been missed due to parsing errors, rules R'_i are generally expected to obtain lower precision and recall (see also Section 5). Ideally, each rule would be associated a weight, and a measure of confidence

1. If a question starts with the verb *Name*:
   R_1 = extract the head of the highest NP immediately after *Name*.
   R'_1 = extract the last word of the maximal sequence of {JJX, NNX, CD} immediately after *Name*.
   Q : *Name the scar-faced bounty hunter of The Old West.*

2. If a question starts or ends with *What/ Which* immediately followed by an NP:
   R_2 = extract the head of the highest NP immediately after the *wh*-word.
   R'_2 = extract the last word of the maximal sequence of {JJX, NNX, CD} immediately after the *wh*-word.
   Q : *What company is the largest Japanese builder?*
   Q : *The corpus callosum is in what part of the body?*

3. If a question starts with *What/ Which/ Who* immediately followed by a BE verb and does not end with a preposition or a past participle verb:
   R_3 = extract the head of the definite highest NP after the BE verb.
   R'_3 = extract the last word of the maximal definite sequence of {JJX, NNX, CD} after the BE verb.
   Q : *What company is the largest Japanese builder?*

4. If a question starts with *What/ Which/ Who*, optionally followed by a non-possessive NP, followed by a NAME verb in passive voice:
   R_4 = extract the head of the highest NP after the NAME verb.
   R'_4 = extract the last word of the maximal sequence of {DT, JJX, NNX, POS, CD} after the NAME verb.
   Q : *What city is sometimes called Gotham?*

5. If a question starts with *What/ Which/ Who*, followed by an interrogative pattern of a NAME verb:
   R_5 = extract the head of the highest NP after the NAME verb.
   R'_5 = extract the last word of the maximal sequence of [DT, JJX, NNX, POS, CD] after the NAME verb.
   Q : *What author did photographer Yousuf Karsh call the shiest man I ever met?*

Fig. 2. Focus Identification Rules and Example Questions.

would be computed for each candidate focus word based on the weights of the rules used to extract it. Such a setting can be obtained by modeling the focus identification task as a binary classification problem in which question words are classified as either part of the question focus or not. Each rule from the rule based approach would give rise to a binary feature whose value would be 1 only for the word positions matched by the rule. The fact that one word may be identified as focus by multiple rules will not be problematic for discriminative learning methods such as Support Vector Machines (SVMs) [11] or Maximum Entropy [12], which can deal efficiently with thousands of overlapping features. For our machine learning approach to focus identification we chose to use SVMs, Question-level features:

1. The question starts with a preposition followed by a *wh*-word.
   Q : *In what U.S. state was the first woman governor elected?*
2. The question starts with *Name*.
   Q : *Name the scar-faced bounty hunter of The Old West.*
3. The question starts with a *wh*-word followed by a BE verb.
4. The question starts with *What/ Which* followed by a BE verb and a bare NP.
   Q : *What are liver enzymes?*
5. The question starts with *What/ Which* followed by a BE verb and ends with a past participle verb, optionally followed by a preposition.
   Q : *What is a female rabbit called?*
6. The question starts with a *wh*-word in an empty WHNP.
   Q : *What are liver enzymes?*
7. The question starts with a *wh*-word followed by an NP.
   Q : *What company is the largest Japanese builder?*
8. The question ends with a preposition.
   Q : *What are Cushman and Wakefield known for?*
9. The first verb after the *wh*-word is not a BE verb.
   Q : *Who killed Gandhi?*
10. The questions starts with ⟨*wh*-word⟩: create a feature for each possible *wh*-word.

Word-level features:

1. Create a feature for each of the rules R_1, R'_1 ... R_5, R'_5.
2. The head of the highest bare NP after the WHNP.
   Q : *What are liver enzymes?*
3. The head of the highest definite NP after the WHNP.
   Q : *What company is the largest Japanese builder?*
4. The head of the highest indefinite NP after the WHNP.
   Q : *What is a cascade?*
5. The head of the highest NP nearest the *wh*-word.
6. The last word of the first maximal sequence of {JJX, NNX, CD} nearest the *wh*-word.
   Q : *What is considered the costliest disaster the insurance industry has ever faced?*
7. The part-of-speech ⟨tag⟩ of the word: create a feature for each possible tag.

Fig. 3. Question-level and Word-level Features .

motivated by their capability to automatically induce implicit features as conjunctions of original features when run with polynomial or Gaussian kernels [13]. The explicit features employed in the SVM model are shown in Figure 3. Apart from the rules R_i and their analogues R'_i from Figure 2, the feature set also contains more atomic features designed to capture simple constraints used in the rules. Splitting rules into their more atomic features means that an isolated error in one part of the parse tree would only affect some of the features, while the rest of the features will still be relevant. Thus, while the rule as a whole may lead to an incorrect extraction, some of its atomic features may still be used to reach a correct decision. Furthermore, if the SVM model is coupled with a polynomial kernel, then more complex parts of the original rules, when seen as conjunctions of elementary features, would be considered as implicit features.

5 Experimental Evaluation

We empirically evaluated the rule based approach and the learning approach on the task of question focus identification using the 2000 manually labeled questions. For the rule based approach, we evaluated 3 rule sets: the set of rules R_1 to R_5, their approximations R'_1 to R'_5, and all the rules from R_1, R'_1 to R_5, R'_5 in a combined set. For the learning approach, we performed 10-fold cross-validation by splitting the dataset into 10 equally sized folds, training on 9 folds and testing on the remaining 1 fold for 10 iterations. To compute the accuracy and F_1 measure, we pooled the results across all 10 folds. We used SVMLight[1] with its default parameters and a quadratic kernel. Table 1 shows the precision, recall, F_1 measure, and accuracy for all four systems. The precision, recall and F_1 measures correspond to the task of focus word extraction and are therefore computed at word level. Accuracy is computed at question level by considering a question correctly classified if and only if the set of focus words found by the system is exactly the same as the set of focus words in the annotation.

Table 1. Experimental Results.

Measure	Rule Based R_1 to R_5	Rule Based R'_1 to R'_5	Rule Based Combined	SVM Quadratic
Precision	93.3%	92.5%	88.1%	**95.2**%
Recall	89.2%	71.6%	90.1%	**91.3**%
F_1	91.2%	80.7%	89.1%	**93.2**%
Accuracy	91.1%	81.8%	88.3%	**93.5**%

As expected, adding the approximation rules to the original rules in the combined system helps by increasing the recall, but hurts the precision significantly. Overall, the learning approach obtains the best performance across all measures, proving that is can exploit useful combinations of overlapping rules and features. Figure 4 shows graphically the precision vs. recall results for the four systems. The curve for the SVM approach was obtained by varying a threshold on the extraction confidence, which was defined to be equal with the distance to the classification hyperplane, as computed by the SVMLight package.

A significant number of errors made by the learning approach are caused by parsing or part-of-speech tagging errors. There are also examples that are misclassified due to the fact that syntactic information alone is sometimes insufficient. For example, questions such as *What is a film starring Jude Law?* have

[1] URL: http://svmlight.joachims.org

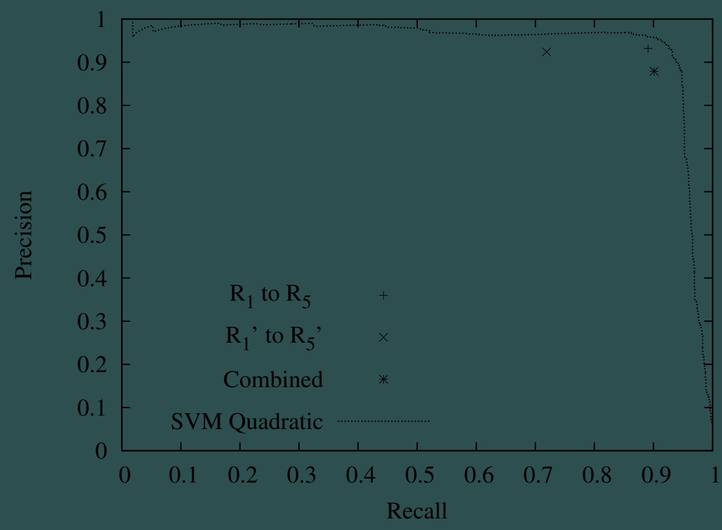

Fig. 4. Precision vs. Recall graphs.

the structure of an implicit definition question, yet they do contain an explicit focus word. The opposite is also possible: for the implicit definition question *What is the "7-minute cigarette"?* the system identifies *cigarette* as focus word. Semantic features that discriminate between proper names and titles may also eliminate some of the errors. The system learns, for example, that words tagged as NNP are unlikely to be focus words, which is why it fails to extract the focus word for the question *Who was President of Afghanistan in 1994?*.

Recently, Mikhailian et al. [14] have proposed a Maximum Entropy approach for identifying the question focus (the *asking point* in their terminology). However, they use the traditional, less comprehensive definition of question focus, whereby a question can have at most one focus noun phrase. In order to empirically compare our learning approach with theirs, we created a second version of the dataset in which the questions were annotated with at most one NP focus. Using exact matching between system output and annotations, our SVM based approach obtains a question level accuracy of 93.7%, which compares favorably with their reported accuracy of 88.8%.

6 Future Work

We plan to augment the SVM model with semantic features, some of them identified in Section 5, in order to further increase the accuracy. We also intend to implement the alternative method mentioned at the beginning of Section 4, in which the identification of question focus is done by classifying the coreference chains extracted from the question as referring to the answer or not.

7 Conclusions

We proposed a comprehensive definition of question focus based on coreference with the answer that eliminates inconsistencies from previous definitions. We designed both a rule based approach and a machine learning approach to question focus identification, and evaluated them on a dataset of 2000 questions manually annotated with focus information. Empirical evaluation of the two approaches shows that focus identification using the new definition can be done with high accuracy, offering the promise of more accurate answer typing models.

References

1. Prager, J.M.: Open-domain question-answering. Foundations and Trends in Information Retrieval **1** (2006) 91–231
2. Ittycheriah, A., Franz, M., Zhu, W.J., Ratnaparkhi, A., Mammone, R.J.: IBM's statistical question answering system. In: Proceedings of the Ninth Text REtrieval Conference, National Institute of Standards and Technology (NIST) (2000)
3. Li, X., Roth, D.: Learning question classifiers. In: Proceedings of the 19th International Conference on Computational linguistics, Taipei, Taiwan, Association for Computational Linguistics (2002) 1–7
4. Pinchak, C., Lin, D.: A probabilistic answer type model. In: Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, The Association for Computer Linguistics (2006)
5. Pinchak, C., Lin, D., Rafiei, D.: Flexible answer typing with discriminative preference ranking. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece (2009) 666–674
6. Steve, A.S., Abney, S., Bacchiani, M., Collins, M., Hindle, D., Pereira, O.: Att at trec-8. In: In Proceedings of the Eighth Text REtrieval Conference (TREC-8. (2000) 317–330
7. Bies, A., Ferguson, M., Katz, K., MacIntyre, R.: Bracketing Guidelines for Treebank II Style Penn Treebank Project. University of Pennsylvania. (1995)
8. Pinchak, C., Bergsma, S.: Automatic answer typing for how-questions. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, New York, Association for Computational Linguistics (2007) 516–523
9. DARPA, ed.: Proceedings of the Seventh Message Understanding Evaluation and Conference (MUC-98), Fairfax, VA, Morgan Kaufmann (1998)
10. Collins, M.: Head-driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania (1999)
11. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons (1998)
12. Berger, A.L., Della Pietra, S.A., Della Pietra, V.J.: A maximum entropy approach to natural language processing. Computational Linguistics **22** (1996) 39–71
13. Schölkopf, B., Smola, A.J.: Learning with kernels - support vector machines, regularization, optimization and beyond. MIT Press, Cambridge, MA (2002)
14. Mikhailian, A., Dalmas, T., Pinchuk, R.: Learning foci for question answering over topic maps. In: Proceedings of the ACL-IJCNLP 2009 Conference, Suntec, Singapore, Association for Computational Linguistics (2009) 325–328