Combining Deep Linguistics Analysis and Surface Pattern Learning: A Hybrid Approach to Chinese Definitional Question Answering

Fuchun Peng, Ralph Weischedel, Ana Licuanan, Jinxi Xu
BBN Technologies
50 Moulton Street, Cambridge, MA, 02138
{fpeng, rweisched, alicuan, jxu}@bbn.com

Abstract

We explore a hybrid approach for Chinese definitional question answering by combining deep linguistic analysis with surface pattern learning. We answer four questions in this study: 1) How helpful are linguistic analysis and pattern learning? 2) What kind of questions can be answered by pattern matching? 3) How much annotation is required for a pattern-based system to achieve good performance? 4) What linguistic features are most useful? Extensive experiments are conducted on biographical questions and other definitional questions. Major findings include: 1) linguistic analysis and pattern learning are complementary; both are required to make a good definitional QA system; 2) pattern matching is very effective in answering biographical questions while less effective for other definitional questions; 3) only a small amount of annotation is required for a pattern learning system to achieve good performance on biographical questions; 4) the most useful linguistic features are copulas and appositives; relations also play an important role; only some propositions convey vital facts.

1 Introduction

Due to the ever increasing large amounts of online textual data, learning from textual data is becoming more and more important. Traditional document retrieval systems return a set of relevant documents and leave the users to locate the specific information they are interested in. Question answering, which combines traditional document retrieval and information extraction, solves this problem directly by returning users the specific answers. Research in textual question answering has made substantial advances in the past few years (Voorhees, 2004).

Most question answering research has been focusing on factoid questions where the goal is to return a list of facts about a concept. Definitional questions, however, remain largely unexplored. Definitional questions differ from factoid questions in that the goal is to return the relevant "answer nuggets" of information about a query. Identifying such answer nuggets requires more advanced language processing techniques. Definitional QA systems are not only interesting as a research challenge. They also have the potential to be a valuable complement to static knowledge sources like encyclopedias. This is because they create definitions dynamically, and thus answer definitional questions about terms which are new or emerging (Blair-Goldensoha et al., 2004).

One success in factoid question answering is pattern based systems, either manually constructed (Soubbotin and Soubbotin, 2002) or machine learned (Cui et al., 2004). However, it is unknown whether such pure pattern based systems work well on definitional questions where answers are more diverse.

Deep linguistic analysis has been found useful in factoid question answering (Moldovan et al., 2002) and has been used for definitional questions (Xu et al., 2004; Harabagiu et al., 2003). Linguistic analy-

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 307–314, Vancouver, October 2005.

sis is useful because full parsing captures long distance dependencies between the answers and the query terms, and provides more information for inference. However, merely linguistic analysis may not be enough. First, current state of the art linguistic analysis such as parsing, co-reference, and relation extraction is still far below human performance. Errors made in this stage will propagate and lower system accuracy. Second, answers to some types of definitional questions may have strong local dependencies that can be better captured by surface patterns. Thus we believe that combining linguistic analysis and pattern learning would be complementary and be beneficial to the whole system.

Work in combining linguistic analysis with patterns include Weischedel et al. (2004) and Jijkoun et al. (2004) where manually constructed patterns are used to augment linguistic features. However, manual pattern construction critically depends on the domain knowledge of the pattern designer and often has low coverage (Jijkoun et al., 2004). Automatic pattern derivation is more appealing (Ravichandran and Hovy, 2002).

In this work, we explore a hybrid approach to combining deep linguistic analysis with automatic pattern learning. We are interested in answering the following four questions for Chinese definitional question answering:

- How helpful are linguistic analysis and pattern learning in definitional question answering?
- If pattern learning is useful, what kind of question can pattern matching answer?
- How much human annotation is required for a pattern based system to achieve reasonable performance?
- If linguistic analysis is helpful, what linguistic features are most useful?

To our knowledge, this is the first formal study of these questions in Chinese definitional QA. To answer these questions, we perform extensive experiments on Chinese TDT4 data (Linguistic Data Consortium, 2002-2003). We separate definitional questions into biographical (Who-is) questions and other definitional (What-is) questions. We annotate some question-answer snippets for pattern learning and we perform deep linguistic analysis including parsing, tagging, name entity recognition, co-reference, and relation detection.

2 A Hybrid Approach to Definitional Question Answering

The architecture of our QA system is shown in Figure 1. Given a question, we first use simple rules to classify it as a "Who-is" or "What-is" question and detect key words. Then we use a HMM-based IR system (Miller et al., 1999) for document retrieval by treating the question keywords as a query. To speed up processing, we only use the top 1000 relevant documents. We then select relevant sentences among the returned relevant documents. A sentence is considered relevant if it contains the query keyword or contains a word that is co-referent to the query term. Coreference is determined using an information extraction engine, SERIF (Ramshaw et al., 2001). We then conduct deep linguistic analysis and pattern matching to extract candidate answers. We rank all candidate answers by predetermined feature ordering. At the same time, we perform redundancy detection based on n-gram overlap.

2.1 Deep Linguistic Analysis

We use SERIF (Ramshaw et al., 2001), a linguistic analysis engine, to perform full parsing, name entity detection, relation detection, and co-reference resolution. We extract the following linguistic features:

1. Copula: a copula is a linking verb such as "*is*" or "*become*". An example of a copula feature is "*Bill Gates is the CEO of Microsoft*". In this case, "*CEO of Microsoft*" will be extracted as an answer to "*Who is Bill Gates?*". To extract copulas, SERIF traverses the parse trees of the sentences and extracts copulas based on rules. In Chinese, the rule for identifying a copula is the POS tag "*VC*", standing for "*Verb Copula*". The only copula verb in Chinese is "是".

2. Apposition: appositions are a pair of noun phrases in which one modifies the other. For example, In "*Tony Blair, the British Prime Minister, ...*", the phrase "*the British Prime Minister*" is in apposition to "*Blair*". Extraction of appositive features is similar to that of copula. SERIF traverses the parse tree and identifies appositives based on rules. A detailed description of the algorithm is documented

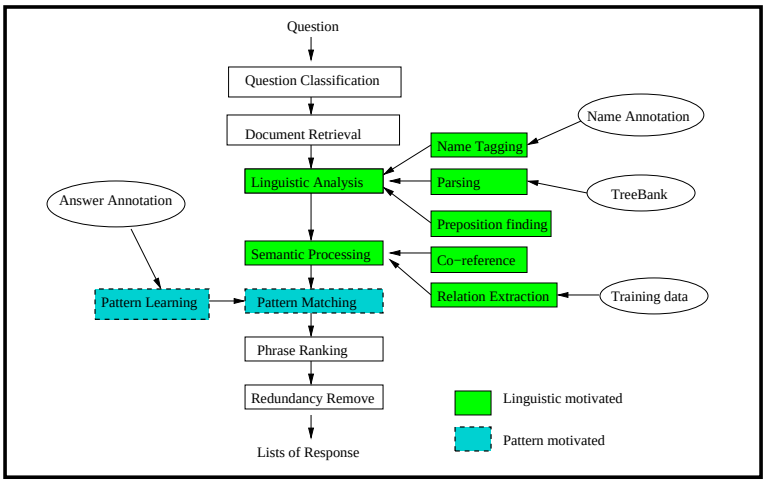

Figure 1: Question answering system structure

in (Ramshaw et al., 2001).

3. Proposition: propositions represent predicate-argument structures and take the form: *predicate(*$role_1$*:* arg_1*, ..,* $role_n$*:* arg_n*)*. The most common roles include logical subject, logical object, and object of a prepositional phrase that modifies the predicate. For example, "Smith went to Spain" is represented as a proposition, *went(logical subject: Smith, PP-to: Spain)*.

4. Relations: The SERIF linguistic analysis engine also extracts relations between two objects. SERIF can extract 24 binary relations defined in the ACE guidelines (Linguistic Data Consortium, 2002), such as spouse-of, staff-of, parent-of, management-of and so forth. Based on question types, we use different relations, as listed in Table 1.

Relations used for Who-Is questions
ROLE/MANAGEMENT, ROLE/GENERAL-STAFF, ROLE/CITIZEN-OF, ROLE/FOUNDER, ROLE/OWNER, AT/RESIDENCE, SOC/SPOUSE, SOC/PARENT, ROLE/MEMBER, SOC/OTHER-PROFESSIONAL
Relation used for What-Is questions
AT/BASED-IN, AT/LOCATED, PART/PART-OF

Table 1: Relations used in our system

Many relevant sentences do not contain the query key words. Instead, they contain words that are co-referent to the query. For example, in "*Yesterday UN Secretary General* **Anan** *Requested Every Side...,* **He** *said ... ".* The pronoun "He" in the second sentence refers to "Anan" in the first sentence. To select such sentences, we conduct co-reference resolution using SERIF.

In addition, SERIF also provides name tagging, identifying 29 types of entity names or descriptions, such as locations, persons, organizations, and diseases.

We also select complete sentences mentioning the term being defined as backup answers if no other features are identified.

The component performance of our linguistic analysis is shown in Table 2.

	Pre.	Recall	F
Parsing	0.813	0.828	0.820
Co-reference	0.920	0.897	0.908
Name-entity detection	0.765	0.753	0.759

Table 2: Linguistic analysis component performance for Chinese

2.2 Surface Pattern Learning

We use two kinds of patterns: manually constructed patterns and automatically derived patterns. A manual pattern is a commonly used linguistic expression that specifies aliases, super/subclass and membership relations of a term (Xu et al., 2004). For example, the expression "*tsunamis, also known as tidal waves*" gives an alternative term for tsunamis. We

use 23 manual patterns for *Who-is* questions and 14 manual patterns for *What-is* questions.

We also classify some special propositions as manual patterns since they are specified by computational linguists. After a proposition is extracted, it is matched against a list of predefined predicates. If it is on the list, it is considered special and will be ranked higher. In total, we designed 22 special propositions for *Who-is* questions, such as 成为(become), 当选为(elected as), and 辞去(resign), 14 for *What-is* questions, such as 位于(located at), 创建于(created at), and 又称为(also known as).

However, it is hard to manually construct such patterns since it largely depends on the knowledge of the pattern designer. Thus, we prefer patterns that can be automatically derived from training data. Some annotators labeled question-answer snippets. Given a query question, the annotators were asked to highlight the strings that can answer the question. Though such a process still requires annotators to have knowledge of what can be answers, it does not require a computational linguist. Our pattern learning procedure is illustrated in Figure 2.

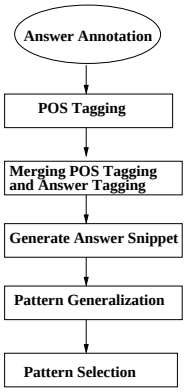

Figure 2: Surface Pattern Learning

Here we give an example to illustrate how pattern learning works. The first step is annotation. An example of Chinese answer annotation with English translation is shown in Figure 3. Question words are assigned the tag *QTERM*, answer words are tagged *ANSWER*, and all other words are assigned *BKGD*, standing for background words (not shown in the example to make the annotation more readable).

To obtain patterns, we conduct full parsing to obtain the full parse tree for a sentence. In our current

Chinese annotation: 到朝鲜进"行破冰之旅"的(美国国务卿 ANSWER)(奥尔布赖特 QTERM), 昨天同朝鲜领袖今正日进行历史性会谈

English translation: (U.S. Secretary of the State ANWER) (Albright QTERM), who visited North Korea for the "ice-breaking trip", had a historical meeting with the leader of North Korea, Kim Jong Il.

Figure 3: Answer annotation example

patterns, we only use POS tagging information, but other higher level information could also be used. The segmented and POS tagged sentence is shown in Figure 4. Each word is assigned a POS tag as defined by the Penn Chinese Treebank guidelines.

(到 P)(朝 鲜 NR)(进 行 VV)(" PU)(破 VV)(冰 NN)(之 旅 NN)(" PU)(的 DEC)(美 国 NR)(国 务 卿 NR)(奥 尔 NR)(布 赖 特 NR)(, PU) (昨 天 NT)(同 DT)(朝鲜 NR)(领袖 NR)(今正日 NN)(举行 VV)(历史性 JJ)(会谈 NN).

Figure 4: POS tagging

Next we combine the POS tags and the answer tags by appending these two tags to create a new tag, as shown in Figure 5.

(到 P/BKGD)(朝 鲜 NR/BKGD)(进 行 VV/BKGD)(" PU/BKGD)(破 VV/BKGD)(冰 NN/BKGD)(之 旅 NN/BKGD)(" PU/BKGD)(的 DEC/BKGD)(美 国 NR/ANSWER)(国 务 卿 NR/ANSWER)(奥 尔 NR/QTERM)(布 赖 特 NR/QTERM)(, PU/BKGD) (昨 天 NT/BKGD)(同 DT/BKGD)(朝 鲜 NR/BKGD)(领 袖 NR/BKGD)(金正日 NN/BKGD)(进行 VV/BKGD)(历史性 JJ/BKGD)(会谈 NN/BKGD)

Figure 5: Combined POS and Answer tagging

We can then obtain an answer snippet from this training sample. Here we obtain the snippet *(美国 国务卿 NR/ANSWER)(TERM)*.

We generalize a pattern using three heuristics (this particular example does not generalize). First, we replace all Chinese sequences longer than 3 characters with their POS tags, under the theory that long sequences are too specific. Second, we also replace NT (time noun, such as 昨天), DT (determiner, such as 这, 那), cardinals (CD, such as 一, 二, 三) and M

(measurement word such as 年月日) with their POS tags. Third, we ignore adjectives.

After obtaining all patterns, we run them on the training data to calculate their precision and recall. We select patterns whose precision is above 0.6 and which fire at least 5 times in training data (parameters are determined with a held out dataset).

3 Experiments

3.1 Data Sets

We produced a list of questions and asked annotators to identify answer snippets from TDT4 data. To produce as many training answer snippets as possible, annotators were asked to label answers exhaustively; that is, the same answer can be labeled multiple times in different places. However, we remove duplicate answers for test questions since we are only interested in unique answers in evaluation.

We separate questions into two types, biographical (Who-is) questions, and other definitional questions (What-is). For "Who-is" questions, we used 204 questions for pattern learning, 10 for parameter tuning and another 42 questions for testing. For "What-is" questions, we used 44 for training and another 44 for testing.

3.2 Evaluation

The TREC question answering evaluation is based on human judgments (Voorhees, 2004). However, such a manual procedure is costly and time consuming. Recently, researchers have started automatic question answering evaluation (Xu et al., 2004; Lin and Demner-Fushman, 2005; Soricut and Brill, 2004). We use Rouge, an automatic evaluation metric that was originally used for summarization evaluation (Lin and Hovy, 2003) and was recently found useful for evaluating definitional question answering (Xu et al., 2004). Rouge is based on n-gram co-occurrence. An n-gram is a sequence of n consecutive Chinese characters.

Given a reference answer R and a system answer S, the Rouge score is defined as follows:

$$Rouge(R,S,N) = \sqrt[N]{\prod_{n=1}^{N} \frac{count_{match}(R,S,n)}{count(R,n)}}$$

where N is the maximum length of n-grams, $count_{match}(R,S,n)$ is the number of common n-grams of R and S, and $count(R,n)$ is the number of n-grams in R. If N is too small, stop words and bi-grams of such words will dominate the score; If N is too large, there will be many questions without answers. We select N to be 3, 4, 5 and 6.

To make scores of different systems comparable, we truncate system output for the same question by the same cutoff length. We score answers truncated at length L times that of the reference answers, where L is set to be 1, 2, and 3. The rationale is that people would like to read at least the same length of the reference answer. On the other hand, since the state of the art system answer is still far from human performance, it is reasonable to produce answers somewhat longer than the references (Xu et al., 2004).

In summary, we run experiments with parameters $N = 3, 4, 5, 6$ and $L = 1, 2, 3$, and take the average over all of the 12 runs.

3.3 Overall Results

We set the pure linguistic analysis based system as the baseline and compare it to other configurations. Table 3 and Table 4 show the results on "*Who-is*" and "*What-is*" questions respectively. The baseline (Run 1) is the result of using pure linguistic features; Run 2 is the result of adding manual patterns to the baseline system; Run 3 is the result of using learned patterns only. Run 4 is the result of adding learned patterns to the baseline system. Run 5 is the result of adding both manual patterns and learned patterns to the system.

The first question we want to answer is how helpful the linguistic analysis and pattern learning are for definitional QA. Comparing Run 1 and 3, we can see that both pure linguistic analysis and pure pattern based systems achieve comparable performance; Combining them together improves performance (Run 4) for "who is" questions, but only slightly for "what is" questions. This indicates that linguistic analysis and pattern learning are complementary to each other, and both are helpful for biographical QA.

The second question we want to answer is what kind of questions can be answered with pattern matching. From these two tables, we can see that patterns are very effective in "*Who-is*" questions while less effective in "*What-is*" questions. Learned patterns improve the baseline from 0.3399

to 0.3860; manual patterns improve the baseline to 0.3657; combining both manual and learned patterns improve it to 0.4026, an improvement of **18.4%** compared to the baseline. However, the effect of patterns on "*What-is*" is smaller, with an improvement of only **3.5%**. However, the baseline performance on "*What-is*" is also much worse than that of "*Who-is*" questions. We will analyze the reasons in Section 4.3. This indicates that answering general definitional questions is much more challenging than answering biographical questions and deserves more research.

Run	Run description	Rouge
(1)	Baseline	0.3399
(2)	(1)+ manual patterns	0.3657
(3)	Learned patterns	0.3549
(4)	(1)+ learned patterns	0.3860
(5)	(2)+ learned patterns	0.4026

Table 3: Results on *Who-is* (Biographical) Questions

Run	Run description	Rouge
(1)	Baseline	0.2126
(2)	(1)+ manual patterns	0.2153
(3)	Learned patterns	0.2117
(4)	(1)+ learned patterns	0.2167
(5)	(2)+ learned patterns	0.2201

Table 4: Results on "*What-is*" (Other Definitional) Questions

4 Analysis

4.1 How much annotation is needed

The third question is how much annotation is needed for a pattern based system to achieve good performance. We run experiments with portions of training data on biographical questions, which produce different number of patterns. Table 5 shows the details of the number of training snippets used and the number of patterns produced and selected. The performance of different system is illustrated in Figure 6. With only 10% of the training data (549 snippets, about two person days of annotation), learned patterns achieve good performance of 0.3285, considering the performance of 0.3399 of a well tuned system with deep linguistic features. Performance saturates with 2742 training snippets (50% training, 10 person days annotation) at a Rouge score of 0.3590, comparable to the performance of a well tuned system with full linguistic features and manual patterns (Run 2 in Table 3). There could even be a slight, insignificant performance decrease with more training data because our sampling is sequential instead of random. Some portions of training data might be more useful than others.

	Training snippets	Patterns learned	Patterns selected
10% train	549	56	33
30% train	1645	144	88
50% train	2742	211	135
70% train	3839	281	183
90% train	4935	343	222
100% train	5483	381	266

Table 5: Number of patterns with different size of training data

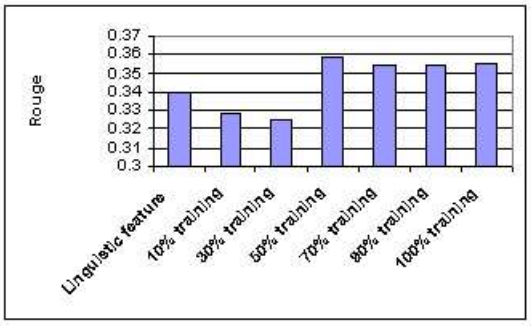

Figure 6: How much annotation is required (measured on biographical questions)

4.2 Contributions of different features

The fourth question we want to answer is: what features are most useful in definitional question answering? To evaluate the contribution of each individual feature, we turn off all other features and test the system on a held out data (10 questions). We calculate the coverage of each feature, measured by Rouge. We also calculate the precision of each feature with the following formula, which is very similar to Rouge except that the denominator here is based on system output $count(S, n)$ instead of reference $count(R, n)$. The notations are the same as

those in Rouge.

$$Precision(R,S,N) = \sqrt[N]{\prod_{n=1}^{N} \frac{count_{match}(R,S,n)}{count(S,n)}}$$

Figure 7 is the precision-recall scatter plot of the features measured on "who is" questions. Interestingly, the learned patterns have the highest coverage and precision. The copula feature has the second highest precision; however, it has the lowest coverage. This is because there are not many copulas in the dataset. Appositive and manual pattern features have the same level of contribution. Surprisingly, the relation feature has a high coverage. This suggests that relations could be more useful if relation detection were more accurate; general propositions are not more useful than whole sentences since almost every sentence has a proposition, and since the high value propositions are identified by the lexical head of the proposition and grouped with the manual patterns.

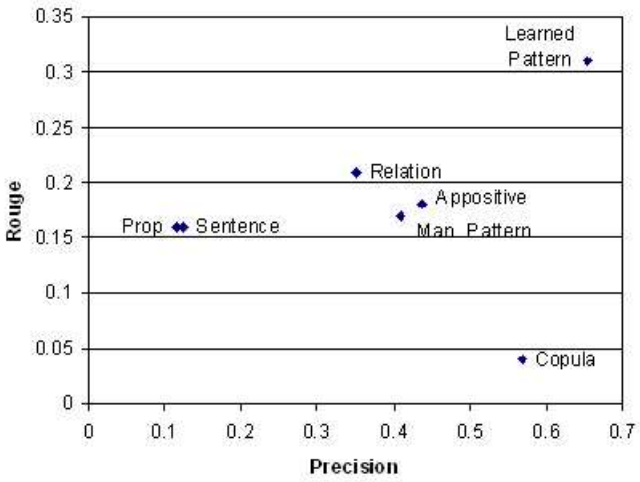

Figure 7: Feature precision recall scatter plot (measured on the biographical questions)

4.3 Who-is versus What-is questions

We have seen that "*What-is*" questions are more challenging than "*Who-is*" questions. We compare the precision and coverage of each feature for "*Who-is*" and "*What-is*" in Table 6 and Table 7. We see that although the precisions of the features are higher for "*What-is*", their coverage is too low. The most useful features for "*What-is*" questions are propositions and raw sentences, which are the worst two features for "*Who-is*". Basically, this means that most of the answers for "*What-is*" are from whole sentences. Neither linguistic analysis nor pattern matching works as efficiently as in biographical questions.

feature	who-is	what-is
copula	0.567	0.797
appositive	0.3460	0.3657
proposition	0.1162	0.1837
relation	0.3509	0.4422
sentence	0.1074	0.1556
learned patterns	0.6542	0.6858

Table 6: Feature Precision Comparison

feature	who-is	what-is
copula	0.055	0.049
appositive	0.2028	0.0026
proposition	0.2101	0.1683
relation	0.2722	0.043
sentence	0.1619	0.1717
learned patterns	0.3517	0.0860

Table 7: Feature Coverage Comparison

To identify the challenges of "*What-is*" questions, we conducted an error analysis. The answers for "*What-is*" are much more diverse and are hard to capture. For example, the reference answers for the question of "什么是国际空间站？ / What is the international space station?" include the weight of the space station, the distance from the space station to the earth, the inner structure of the space station, and the cost of its construction. Such attributes are hard to capture with patterns, and they do not contain any of the useful linguistic features we currently have (copula, appositive, proposition, relation). Identifying more useful features for such answers remains for future work.

5 Related Work

Ravichandran and Hovy (2002) presents a method that learns patterns from online data using some seed questions and answer anchors. The advantage is that it does not require human annotation. However, it only works for certain types of questions that

have fixed anchors, such as "where was X born". For general definitional questions, we do not know what the anchors should be. Thus we prefer using small amounts of human annotation to derive patterns. Cui et al. (2004) uses a similar approach for unsupervised pattern learning and generalization to soft pattern matching. However, the method is actually used for sentence selection rather than answer snippet selection. Combining information extraction with surface patterns has also seen some success. Jikoun et al. (2004) shows that information extraction can help improve the recall of a pattern based system. Xu et al. (2004) also shows that manually constructed patterns are very important in answering English definitional questions. Hildebrandt et al. (2004) uses manual surface patterns for target extraction to augment database and dictionary lookup. Blair-Goldensohn et al. (2004) apply supervised learning for definitional predicates and then apply summarization methods for question answering.

6 Conclusions and Future Work

We have explored a hybrid approach for definitional question answering by combining deep linguistic analysis and surface pattern learning. For the first time, we have answered four questions regarding Chinese definitional QA: deep linguistic analysis and automatic pattern learning are complementary and may be combined; patterns are powerful in answering biographical questions; only a small amount of annotation (2 days) is required to obtain good performance in a biographical QA system; copulas and appositions are the most useful linguistic features; relation extraction also helps.

Answering "*What-is*" questions is more challenging than answering "*Who-is*" questions. To improve the performance on "*What-is*" questions, we could divide "*What-is*" questions into finer classes such as organization, location, disease, and general substance, and process them specifically.

Our current pattern matching is based on simple POS tagging which captures only limited syntactic information. We generalize words to their corresponding POS tags. Another possible improvement is to generalize using automatically derived word clusters, which provide semantic information.

Acknowledgements This material is based upon work supported by the Advanced Research and Development Activity (ARDA) under Contract No. NBCHC040039. We are grateful to Linnea Micciulla for proof reading and three anonymous reviewers for suggestions on improving the paper.

References

S. Blair-Goldensoha, K. McKeown, and A. Hazen Schlaikjer. 2004. Answering Definitional Questions: A Hybrid Approach. *New Directions In Question Answering.*, pages 47–58.

H. Cui, M. Kan, and T. Chua. 2004. Unsupervised Learning of Soft Patterns for Definitional Question Answering. In *WWW 2004*, pages 90–99.

S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, and J. Bensley. 2003. Answer Mining by Combining Extraction Techniques with Abductive Reasoning. In *TREC2003 Proceedings*.

W. Hildebrandt, B. Katz, and J. Lin. 2004. Answering Definition Questions with Multiple Knowledge Sources. In *HLT-NAACL 2004*, pages 49–56.

V. Jijkoun, M. Rijke, and J. Mur. 2004. Information Extraction for Question Answering: Improving Recall Through Syntactic Patterns. In *COLING 2004*.

J. Lin and D. Demner-Fushman. 2005. Automatically Evaluating Answers to Definition Questions. In *ACL2005*. to appear.

C. Lin and E. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *HLT-NAACL 2003*.

D. Miller, T. Leek, and R. Schwartz. 1999. A Hidden Markov Model Information Retrieval System. In *SIGIR 1999*, pages 214 – 221.

D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu. 2002. Performance Issues and Error Analysis in an Open-Domain Question Answering System. In *ACL2002*.

L. Ramshaw, E. Boshee, S. Bautus, S. Miller, R. Stone, R. Weischedel, and A. Zamanian. 2001. Experiments in Multi-Model Automatic Content Extraction. In *HLT2001*.

D. Ravichandran and E. Hovy. 2002. Learning surface text patterns for a Question Answering System. In *ACL2002*, pages 41–47.

R. Soricut and E. Brill. 2004. A Unified Framework For Automatic Evaluation Using N-Gram Co-occurrence Statistics. In *ACL 2004*, pages 613–620.

M. Soubbotin and S. Soubbotin. 2002. Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach. In *TREC2002 Proceedings*.

E. Voorhees. 2004. Overview of the TREC 2003 Question Answering Track. In *TREC Proceedings*.

R. Weischedel, J. Xu, and A. Licuanan. 2004. A Hybrid Approach to Answering Biographical Questions. *New Directions In Question Answering.*, pages 59–70.

J. Xu, R. Weischedel, and A. Licuanan. 2004. Evaluation of an Extraction-based Approach to Answering Definitional Questions. In *SIGIR 2004*, pages 418–424.