A SURVEY ON QUESTION CLASSIFICATION TECHNIQUES FOR QUESTION ANSWERING

Natsuda Laokulrat

School of Engineering, The University of Tokyo

Emails: natsuda@logos.t.u-tokyo.ac.jp

ABSTRACT

Question answering is one of the oldest and challenging tasks in natural language processing. The goal is to build systems that are able to automatically answer questions posed by human in a natural language. This survey focuses on question classification task, which is a subtask in question answering. Question classification aims to associate a category to each question, typically representing the semantic class of its answer. It is of major importance in the question answering process, since it is the basis of several key decisions. This survey presents basic introduction about question answering, some approaches for question classification, including rule-based approach and machine learning approach, and the accuracy comparison among them.

Index Terms – Natural Language Processing; NLP; Machine learning; Question answering; Information retrieval

1. INTRODUCTION

Question answering (QA) is related to natural language processing (NLP) and information retrieval. The goal is to build systems that are able to automatically answer questions posed by human in a natural language. There are some well-known applications that succeeded in QA task such as Apple's Siri, IBM Watson, Google Search, and WolframAlpha.

Apple's Siri, an application for Apple's iOS, uses a natural language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Web services. Figure 1 demonstrates how Siri answers the questions.

IBM Watson is an artificial intelligence computer system capable of answering questions posed in natural language, developed in IBM's DeepQA project (www-03.ibm.com/innovation/us/watson/index.html). In 2011, Watson competed on the quiz show Jeopardy!, as shown in Figure 2, and received the first prize. It had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage, including the full text of Wikipedia, but was not connected to the Internet during the game.

Figure 1. Apple's Siri

Figure 2. IBM Watson

Google Search also has the ability to answer factoid questions, as shown in Figure 3. Another state-of-the-art question answering system is WolframAlpha (www.wolframalpha.com), developed by Wolfram Research. It can give answers on facts and data and calculates answers across a range of topics, including science, nutrition, history, geography, engineering, mathematics, linguistics, sports, etc.

Question classification, a QA subtask, aims to associate a category to each question, typically representing the semantic class of its answer. This step is of major importance in the QA process, since it is the basis of several key decisions. For instance, classification helps reducing the number of possible answer candidates, as only answers

Figure 3. Google Search

matching the question category should be taken into account.

The remaining of the paper is organized as follows: Section 2 gives an overall explanation about question answering system. Section 3 describes some basic knowledge about question types and question classification. Section 4, 5, and 6 describe the current approaches for question classification. Section 7 is the conclusion of this report.

2. QUESTION ANSWERING SYSTEM

Question answering has 3 main phases, as described in [1], including question processing, passage retrieval, and answer processing. Fig Figure 4 illustrates the flow of the QA system.

2.1. Question processing

The goal of this phase is to extract a keyword query and an answer type from the question. The keyword query is used as an input to an IR system. The answer type is used for specifying the kind of entity that would constitute a reasonable answer to the question.

The question classification task is to classify the question by its expected answer type. If the answer type is known for a question, we can avoid looking at every sentence or noun phase in the

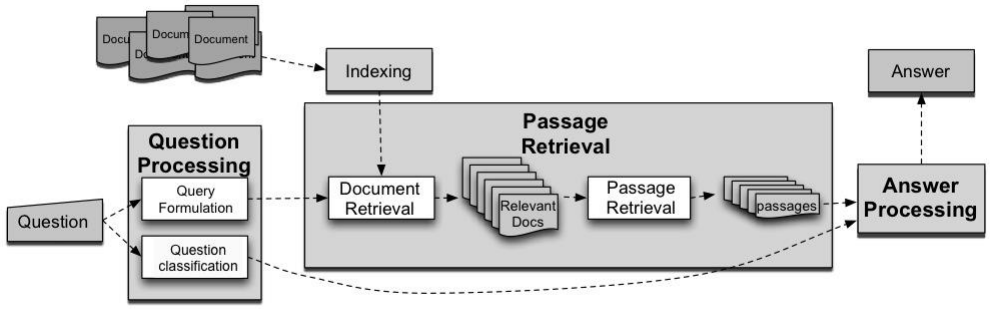

Figure 4. (from [1]) Question answering system

entire set of documents for an answer. For example, a question like *"Who founded Google?"* expects an answer of type PERSON. A question like *"Which country has the largest population?''* expects an answer of type COUNTRY.

Knowing an answer type is also important for presenting the answer. For instance, a DEFINITION question like *"What is a prism?"* might use a simple answer template like *"A prism is ..."*, whereas an answer to a BIOGRAPHY question like *"Who is Barack Obama?"* might use a biography-specific template for presenting the an answer.

2.2. Passage retrieval

The query that was created in the question-processing phase is then used to query an information-retrieval system. The result of this document retrieval stage is a set of documents.

The next stage is to extract a set of potential answer passages in the set of documents. The remaining passages are then ranked; either by handwritten rules or by supervised training with machine learning techniques.

2.3. Answer processing

The final stage is to extract a specific answer from the passage so as to be able to present the use with an answer. In the answer pattern extraction for answer processing, we use the information about expected answer type together with regular expression patterns. For example, for questions with a HUMAN answer type, we run the answer type or named entity tagger on the candidate passage or sentence and return whatever entity with type HUMAN.

For some questions, instead of using answer types, we use handwriting regular expression patterns to help extract the answers. Some examples are shown in Figure 9.

3. QUESTION CLASSIFICATION

The goal of question classification is to map a question into a category that represents the type of information that is expected to be present in the final answer. Question classification is a very important step in the question answering process, as the selected question category can be used for different purposes. The importance of question classification is listed below [5].

Coarse	Fine
ABBREVIATION	Abbreviation, expansion
DESCRIPTION	Definition, description, manner, reason
ENTITY	Animal, body, color, creative, currency, medical disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUMAN	Description, group, individual, title
LOCATION	City, country, mountain, other, state
NUMERIC	Code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

Figure 5. Two-layer question taxonomy

- It can help narrowing down the number of possible candidate by using the selected question category as a matching criterion to filter out candidate answers that look likely.
- Depending on the question category, different strategy can be chosen to find an answer. For example, if a question is classified in to the category DEFINITION, possible answers can be searched in encyclopedic sources, such as Wikipedia.
- Misclassified question can hinder the ability to reach a correct answer, because it can lead to wrong assumptions about the question.

3.1. Types of question

Question can be roughly divided into 2 types: factoid and non-factoid questions. We call them factoid questions if the information is a simple fact, and particularly if this fact has to do with a named entity like a person organization, or location, such as

- Where is Louvre Museum located?
- How many calories are there in two slices of apple pie?
- What currency is used in China?

Non-factoid questions are complex or narrative questions such as

Tag	Example
ABBREVIATION	
abbreviation	Whats the abbreviation for limited partnership?
expansion	What does the "c" stand for in the equation E = mc2?
DESCRIPTION	
definition	What are tannins?
reason	What caused the Titanic to sink?
ENTITY	
currency	What currency is used in China?
word	What's the singular of dice?
HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
LOCATION	
city	What's the oldest capital city in the Americas?
mountain	What is the highest peak in Africa?
NUMERIC	
size	What is the size of Argentina?
date	What is the date of Boxing Day?

Figure 6. Classification example

- What do scholars think about Jefferson's position on dealing with pirates?}
- In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?}

3.2. Question taxonomy

Figure 5 shows the two-layered question taxonomy proposed in [2], which contains 6 coarse grained categories and 50 fine-grained categories. This taxonomy is widely used in the machine learning community.

Figure 6 shows how the taxonomy in Figure 5 is used in question classification task.

4. RULE-BASED APPROACH

Question classifier can be built by handwriting rules. The Webclopedia QA Typology [3][4], for example, contains 276 hand-written rules associated with the approximately 180 answer types.

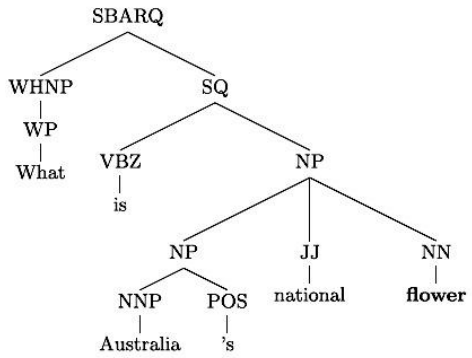

Figure 7. Parse tree. The headword is in bold face.

Parent	Directory	Priority list
S	Left	VP S FRAG SBAR ADJP
SBARQ	Left	SQ S SINV SBARQ FRAG
SQ	Left	NP VP SQ
NP	Right by position	NP NN NNP NNPS NNS NX
PP	Left	WHNP NP WHADVP SBAR
WHNP	Left	NP
WHPP	Right	WHNP WHADVP NP SBAR

Figure 8. Subset of head-rules to determine the question headwords

4.1. Handwritten rules

A regular expression rule can be used for detecting an answer type. For instance, the regular expression for detecting an answer type like BIOGRAPHY might be "*Who { is | was | are | were} PERSON*".

A set of patterns from [4] introduced by Hovy et al. for detecting a QA type of a person name is shown in Figure 9 where the answer templates are also given.

Silva et al. [5] also built a rule-based classifier as shown in Figure 10. The rule-based classifier starts by triggering a set of 60 manually built patterns introduced in Subsection 3.2, that are matched against each question. If the match is successful, a category is returned and the question is classified; otherwise the classifier searches for the question headword and extracts it. Then, the headword hypernyms are followed until one is associated with a possible question category. For instance, the manually built patterns are able to correctly classify the sentence "*When did Hawaii become a state?*" with the fine-grained category NUMERIC:DATE. However, no pattern matches the question "*What person's head is on a dime?*". Therefore, its headword -- person -- is (correctly) identified. By following its hypernyms, the classifier correctly tags it as HUMAN:INDIVIDUAL.

4.2. Headword Extraction

If the match fails, the classifier will search for the question headword and extract it. The headword of a given question is a word that represents the object that is being sought after. It serves as an important clue to the question's category. For example,

- What is Australia's national **flower**?
- Which **country** are Godiva chocolates from?

In the first example, the headword ***flower*** provides the classifier with an important clue to correctly classify the question to ENTITY:PLANT.

A rule-based method is used for headword extraction, allowing a precise headword extraction. The method attained an accuracy of 96.9% for coarse-grained categories. This approach for the extraction of the question headword requires a parse tree of the question. The parse tree for the first example question is displayed in Figure 7.

The set of rules, the head rules, which partially shown in Figure 8 is then applied to the resulting parse tree in order to decide which node is the head or contains it. This process is then repeated recursively until a terminal node is reached. The

Question examples	Question templates
Who was Johnny Mathis' high school track coach?	who be ⟨entity⟩'s ⟨role⟩
Who was Lincoln's Secretary of State?	
Who was President of Turkmenistan in 1994?	who be ⟨role⟩ of ⟨entity⟩
Who is the composer of Eugene Onegin?	
Who is the CEO of General Electric?	
Actual answers	**Answer templates**
Lou Vasquez, track coach ofand Johnny Mathis	⟨person⟩, ⟨role⟩ of ⟨entity⟩
Signed Saparmurad Turkmenbachy [Niyazov], president of Turkmenistan	⟨person⟩ ⟨role-title*⟩ of ⟨entity⟩
...Turkmenistans President Saparmurad Niyazov	⟨entity⟩s ⟨role⟩ ⟨person⟩
...in Tchaikovsky's Eugene Onegin...	⟨person⟩'s ⟨entity⟩
Mr. Jack Welch, GE chairman...	⟨role-title⟩ ⟨person⟩ ... ⟨entity⟩ ⟨role⟩
...Chairman John Welch said ...GE's	⟨subject⟩|⟨psv object⟩ of related role-verb

Figure 9. Handwritten rules

Category	Question pattern description	Example
Abbrev.:Expansion	Begins with What do(es) and ends with an acronym - i.e., a sequence of capital letters possibly intervened by dots -, followed by stands for/mean;	What does AIDS mean?
	Begins with What is/are and ends with an acronym	What is F.B.I.?
Description:Def.	Begins with What is/are and is followed by an optional determiner and a sequence of nouns	What is ethology?
Entity:Term	Begins with What do you call	What do you call. . .
Entity:Substance	Begins with What is/are and ends with composed/made of What is glass made of?	
Description:Reason	Begins with What causes What causes asthma?	
Human:Description	Begins with Who is/was and is followed by a proper noun	Who was Mozart?

Figure 10. A set of question pattern used to avoid extracting a headword, when not needed

head-rules used in this work are a heavily modified version of those given in [6], specifically tailored to extract headwords from questions.

The question category is then obtained from the headword, by using WordNet (Fellbaum 1998), a lexical database for the English language. Consider the following examples.

- *What **explorer** was nicknamed Iberia's Pilot?*
- *What **actor** first portrayed James Bond?*
- *What **dictator** has the nickname "El Maximo"?*

Even though all of the above examples fall under the HUMAN:INDIVIDUAL category, the question headword is different in all of them, which limits the usefulness of the headword in the question classification process. These headwords do, however, share a common trait: they are all subordinates (hyponyms) of the word person, that is to say, they are all more specific senses of person, the superordinate (hypernym). Knowing this information would be useful to accurately classify the previous questions. Silva et al. [5] exploited this observation by using WordNet's lexical hierarchy to associate the headword with a higher-level semantic concept, which represents a question category. Some of these clusters are shown in Figure 11.

Category (Cluster name)	Synsets	Example hyponyms
Entity:Animal	Animal, animate_being, beast, brute, creature, fauna	Mammal, fish, cat
	animal_group	Flock, herd, breed
Entity:Creative	Show	Movie, film, tv show
	Music	Song, tune, hymn
	Writing, written material, piece of writing	Book, poem, novel
Entity:Plant	Vegetation, flora, botany	Forest, garden
	Plant, flora, plant life	Flower, tree, shrub
Human:Individual	Person, individual, someone, somebody, mortal	Actor, leader
	Spiritual being, supernatural being	God, angel, spirit
	Homo, man, human being, human	Homo sapiens
Numeric:Distance	Distance	Altitude, elevation
	Dimension	Width, length, height

Figure 11. Examples of clusters that aggregate similar synsets together

5. MACHINE-LEARNING-BASED APPROACH

Most modern question classifiers nowadays are based on supervised machine learning techniques. These classifiers are trained on databases of questions that have been hand-labeled with an answer type such as the taxonomy introduced in Subsection 3.2.

In the rest of this section, some types of features that were used for training the classifier are introduced.

5.1. Bag-of-words

The most common approach to question classification is bag-of-words (BoW). BoW is a representation of text as an unordered collection of words, disregarding grammar and even word order. Consider these 2 questions

- Question 1: Who is the tallest man in the world?
- Question 2: Who is the tallest man in Japan?

Based on the 2 sentences, a dictionary is constructed as: {who:1, is:2, the:3, tallest:4, man:5, in:6, world:7, japan:8} which has 8 distinct words. By using the indexes of the dictionary, each sentence is represented by a 8-entry vector:

- Question 1: [1, 1, 2, 1, 1, 1, 1, 0]
- Question 2: [1, 1, 1, 1, 1, 1, 0, 1]

where each entry of the vectors refers to count of the corresponding entry in the dictionary.

5.2. Bag-of-ngrams

An n-gram is a consecutive sequence of n items from a given sequence of text (sentence). An n-gram of size 1 is referred to as a ``unigram'', size 2 is a ``bigram'', and size 3 is a "trigram". Consider the question: *"Who is the tallest man in the world?"*

The bigram representations of the question is {(who, is), (is, the), (the, tallest), (tallest, man), (man, in), (in, the), (the, world)}.

The trigram representations of the question is {(who, is, the), (is, the, tallest), (the, tallest, man), (tallest, man, in), (man, in, the), (in, the, world)}.

For example, the bigram dictionary for the two sentences in Subsection 5.1 is {(who, is):1, (is, the):2, (the, tallest):3, (tallest, man):4, (man, in):5, (in, the):6, (the, world):7, (in, japan):8}.

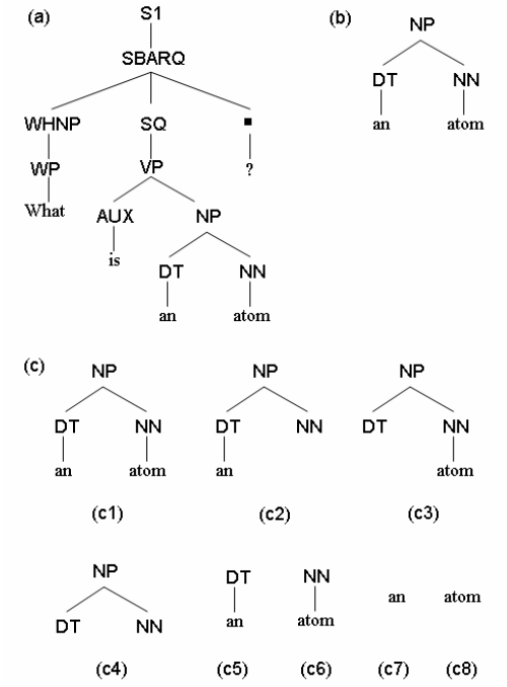

Figure 12. (a) The syntactic tree of the sample question. (b) One of the sub-trees of a. (c) All tree fragments of a within the extent of b.

5.3. Syntactic parse tree

There might be a limit imposed by the representation of questions, which ignores syntax, so including syntactic information might be helpful. For example, the two questions *"Which university did the president graduate from?"* and *"Which president is a graduate of the Harvard University?"* could be discriminated by their different syntactic structures, while the BoW approach can hardly distinguish them.

Zhang and Lee [7] proposed to use a special kernel function called tree kernel to enable the Support Vector Machine (SVM) to take advantage of the syntactic structures of questions.

Figure 12 shows the syntactic tree and tree fragments for the question *"What is an atom?"*.

tree fragment	size	depth
c1	3	4
c2	2	4
c3	2	4
c4	2	4
c5	1	5
c6	1	5
c7	0	6
c8	0	6

Figure 13. The size & depth of each tree fragment in Figure 12

Figure 13 displays the size and depth of each tree fragment in Figure 12. The tree fragments of a syntactic tree are all its sub-trees, which include at least one terminal symbol (word) or one production rule, with the restriction that no production rules can be broken into incomplete parts.

5.3.1. Tree kernel

A key property of the SVM is that the only operation it requires is the computation of dot products between pairs of examples. Collins et al. [8] proposed this kernel method to convert sentences into feature vectors to use as input to SVM, with an attempt to capture considerably more structural information by considering all tree fragments that occur in a parse tree. Suppose that we have 2 syntactic parse trees. Roughly, the kernel counts the number of tree fragments that occur in both syntactic parse trees. Zhang and Lee [7] proposed to weight them with size and depth of each tree fragment. Please refer to the mathematical model in [7].

5.3.2. Shallow semantic parse tree

Bloehdorn et al. [9] claimed that shallow semantic representations, bearing more compact

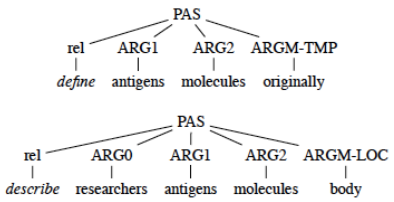

Figure 14. Shallow semantic tree

Query	Results	Normalized
President is a person	259	0.8662
President is a place	9	0.0301
President is an organization	11	0.0368
President is a measure	20	0.0669
President is a date	0	0

Figure 15. Example of using the Internet to extract features for question classification

information, could prevent the sparseness of deep structural approaches (syntactic parse tree) and the weakness of BOW models. They applied Semantic Role Labeling (SRL) [10] to QA system. By using PropBank(PB) [11], the Penn English Treebank with the addition of semantic information, SRL tasks can be done accurately. The goal is to label syntactic nodes with specific argument labels that preserve the similarity of roles such as *the window* in *John broke the window* and *the window broke*.

Consider the PB annotation: [ARG1 Antigens] were [AM-TMP originally] [rel defined] [ARG2 as non-self molecules]. Such annotation can be used to design a shallow semantic representation that can be matched against other semantically similar sentences, e.g. [ARG0 Researchers] [rel describe] [ARG1 antigens] [ARG2 as foreign molecules] [ARGM-LOC in the body]. They are represented in Predicate-Argument Structures (PAS) as shown in Figure 14.

The tree kernel explained in Subsection 5.3 is then applied to the PAS to be used as input to SVM.

5.3.3. Language independent method

The aforementioned approaches have the disadvantage of being targeted to a particular language. Solorio et al. [12] presented a simple approach that exploits lexical features and the Internet to train a SVM classifier. The main feature of this method is that it can be applied to different languages without requiring major modifications.

The procedure for gathering the information from the web is as follows: a set of heuristics is used to extract from the question a word *w*, or set of words, that will complement the queries submitted for the search. Then a search engine is used, in this case Google, and queries are submitted using the word *w* in combination with all the possible semantic classes. For instance, for the question *"Who is the President of the French Republic?"* the word President is extracted using heuristics, and then 5 queries are run in the search engine, one for each possible class. These queries take the following form:

- President is a person
- President is a place
- President is a date
- President is a measure
- President is an organization

The number of results returned by Google for each query is counted and normalized, as displayed in Figure 15. The resultant numbers are the values for the attributes used by the learning algorithm.

	H	C	H+C	U+H	U+C	U+H+C
Coarse	63.2%	92.0%	94.6%	91.8%	95.0%	95.0%
Fine	39.0%	83.4%	88.8%	84.2%	90.2%	90.8%

Figure 16. Accuracy of the hybrid classifier for coarse- and fine-grained categories

Approach	Coarse	Fine
Handwritten rules	87.0%	83.2%
Bag-of-words	85.8%	80.2%
Bag-of-ngrams	87.4%	79.2%
Syntactic parse tree	90.0%	80.2%
Hybrid	95.0%	90.8%

Figure 17. Accuracy of each classifier for coarse- and fine-grained categories

6. HYBRID APPROACH

Silva et al. [5] also proposed the hybrid approach. The information provided by the rule-based classifier - both headwords (H) and categories (C) - is used to generate the feature set for training, training, and merged with the information provided by the question unigrams (U).

It is when these features are combined with unigrams that the classifier holds the best results: an increase of 8.0 $¥%$ and 7.6 $¥%$ compared with the rule-based classifier, for coarse-grained and fine-grained categories, respectively. The results are shown in Figure 16.

7. CONCLUSION

Basic introduction on QA system has been introduced along with the example applications. Some approaches for question classification, including rule-based, machine-learning-based, and hybrid approaches, have been reported here. From the survey, the best one can achieve 95.0$% on coarse-grained category and 90.8% on fine-grained category.

The results of some of the mentioned classifiers are summarized in Figure 17. Note that, all of them used the two-layer taxonomy introduced in 3.2 and SVM classifiers are used. The results from shallow semantic structure and internet-based approach are not shown in Figure 17 because they are not comparable to others.

REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2009.

[2] X. Li and D. Roth, "Learning question classifiers," in *Proc. 19th Int. Conf. Computational linguistics - Volume 1*, ser. COLING'02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1-7.

[3] E. Hovy *et al.*, "Question answering in webclopedia," in *Proc. 9th Text REtrieval Conf. (TREC-9)*, 2000, pp. 655-664.

[4] E. Hovy *et al.*, "A question/answer typology with surface text patterns," in *Proc. second Int. conf. Human Language Technology Research*, ser. HLT'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 247-251.

[5] J. Silva *et al.*, "From symbolic to sub-symbolic information in question classification,"

Artificial Intelligence Review, vol. 35, pp. 137-154, 2011, 10.1007/s10462-010-9188-4.

[6] M. Collins, “Head-driven statistical models for natural language parsing," *Comput. Linguist.*, vol. 29, no. 4, pp. 589-637, Dec. 2003.

[7] D. Zhang and W. S. Lee, “Question classification using support vector machines," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Research and development in information retrieval*, ser. SIGIR'03. New York, NY, USA: ACM, 2003, pp. 26-32.

[8] M. Collins and N. Duffy, “Convolution kernels for natural language," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 625-632.

[9] S. Bloehdorn and A. Moschitti, “Exploiting structure and semantics for expressive text kernels," in *Proc. Sixteenth ACM Conf. Information and Knowledge Management, CIKM 2007*, Lisbon, Portugal, November 2007, pp. 861-864.

[10] L. Marquez *et al.*, “A robust combination strategy for semantic role labeling," in *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT'05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 644-651.

[11] P. Kingsbury and M. Palmer, “From treebank to propbank," in *Proc. 3rd Int. Conf. Language Resources and Evaluation (LREC-2002)*, 2002.

[12] T. Solorio *et al.*, “A language independent method for question classification," in *Proc. 20th Int. Conf. Computational Linguistics*, ser. COLING'04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.