Answer Validation by Information Distance Calculation

Fangtao Li, Xian Zhang, Xiaoyan Zhu
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
zxy-dcs@mail.tsinghua.edu.cn

Abstract

In this paper,an information distance based approach is proposed to perform answer validation for question answering system. To validate an answer candidate, the approach calculates the conditional information distance between the question focus and the candidate under certain condition pattern set. Heuristic methods are designed to extract question focus and generate proper condition patterns from question. General search engines are employed to estimate the Kolmogorov complexity, hence the information distance. Experimental results show that our approach is stable and flexible, and outperforms traditional *tfidf* methods.

1 Introduction

Question answering(QA) system aims at finding exact answers to a natural language question. In order to correctly answer a question, several components are implemented including question classification, passage retrieval, answer candidates generation, answer validation etc. Answer Validation is to decide whether the candidate answers are correct or not, or even to determine the accurate confidence score to them. Most of QA systems employ answer validation as the last step to identify the correct answer. If this component fails, it is impossible to enable the question to be correctly answered.

Automatic techniques for answer validation are of great interest among question answering research. With automatic answer validation, the system will carry out different refinements of its searching criteria to check the relevance of new candidate answers. In addition, since most of QA systems rely on complex architectures and the evaluation of their performances requires a huge amount of work, the automatic assessment of candidates with respect to a given question will speed up both algorithm refinement and testing.

Currently, answer validation is mainly viewed as a classification problem or ranking problem. Different models, such as Support Vector Machine (Shen and Klakow, 2006) and Maximum Entropy Model (Ittycheriah et al., 2001), are used to integrate sophisticated linguistic features to determine the correctness of candidates. The answer validation exercise (Penas et al. , 2007) aims at developing systems able to decide whether the answer is correct or not. They formulate answer validation as a text entailment problem. These approaches are dependent on sophisticated linguistic analysis of syntactic and semantic relations between question and candidates. It is quite expensive to use deep analysis for automatic answer validation, especially in large scale data set. Thus it is appropriate to find an alternative solution to this problem. Here, we just consider the English answer validation task.

This paper proposes a novel approach based on information retrieval on the Web. The answer validation problem is reformulated as distance calculation from an answer candidate to a question. The hypothesis is that, among all candidates, the correct answer has the smallest distance from question. We employ conditional normalized min distance, which is based on Kolmogorov Complexity theory (Li and Vitanyi, 1997), for this task. The distance measures the relevance between question

Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering (IR4QA), pages 42–49
Manchester, UK. August 2008

focus and candidates conditioned on a surface pattern set. For distance calculation, we first extract the question focus, and then a hierarchical pattern set is automatically constructed as condition. Since Kolmogrov Complexity can be approximated through frequency counts. Two types of search engine “Google” and “Altavista” are used to approximate the distance.

The paper is organized as follows: Section 2 describes related work. The fundamental Kolmogorov Complexity theory is introduced in Section 3. Section 4 presents our proposed answer validation method based on information retrieval. In Section 5, we describe the experiments and discussions. The paper is concluded in Section 6.

2 Related Work

Answer Validation is an emerging topic in Question Answering, where open domain systems are often required to rank huge amounts of answer candidates. This task can be viewed as a classification problem or re-ranking problem.

Early question answering systems focused on employing surface text patterns (Subbotin and Subbotin, 2001) for answer validation. Xu et al. (2003) identified that pattern-based approaches got bad performances due to poor system recall. Some researchers exploited machine learning techniques with rich syntactic or semantic features to measure the similarity between question and answer. Ittycheriah et al. (2001) used Maximum Entropy model to combine rich features and automatically learn feature weights. These features included query expansion features, focus features, named entity features, dependency relation features, pattern features et al. Shen and Klakow (2006) presented three methods, including feature vector, string kernel and tree kernel, to represent surface text features and parse tree features in Support Vector Machines. Ko et al. (2007) proposed a probabilistic graphical model to estimate the probability of correctness for all candidate answers. Four types of features were employed, including knowledge-based features, data-driven features, string distance feature and synonym features.

Started in 2006, the annual Answer Validation Exercise (Penas et al. , 2007) aims to develop systems to decide if the answer to a question is correct or not. The English answer validation task is reformulated as a Text Entailment problem. The triplet, including question, answer and supporting text, is given. The system determines if the supporting text can entail the hypothesis, which is a reformulation from the question and answer. All participants used lexical processing, including lemmatization and part-of speech tagging. Some systems used first order logic representations, performed semantic analysis and took the validation decision with a theorem proof.

The above approaches should process deep syntactic and semantic analysis for either questions or candidate answers. The annotated linguistic resource is hard to acquire for the supervised classification problem. Another alternative solution for answer validation is to exploit the redundancy of large scale data. Eric et al. (2007) developed AskMSR question answering system. They focus on the Web as a gigantic data repository with tremendous redundancy that can be exploited to extract the correct answer. Lin (2007) implemented another Web-based question answering system, named ARANEA, which is used approximate tfidf method for answer validation.

3 Preliminaries

3.1 Kolmogorov complexity

Kolmogorov complexity , or *algorithm entropy* , $K(x)$ of a string x is the length of the shortest binary program to compute x. It defines randomness of an individual string. Kolmogorov complexity has been widely accepted as an information theory for individual objects parallel to that of Shannon's information theory which is defined on an ensemble of objects. It has also found many applications in computer science such as average case analysis of algorithms (Li and Vitanyi, 1997). For a universal Turing machine U, the Kolmogorov complexity of a binary string x condition to another binary string y, $K_U(x|y)$, is the length of the shortest (prefix-free) program for U that outputs x with input y. It has been proved that for different universal Turing machine U', for all x, y

$$K_U(x|y) = K_{U'}(x|y) + C,$$

where the constant C depends only on U'. Thus we simply write $K_U(x|y)$ as $K(x|y)$. Define $K(x) = K(x|\epsilon)$, where ϵ is the empty string. For formal definitions and a comprehensive study of Kolmogorov complexity, see (Li and Vitanyi, 1997).

3.2 Information Distance

Based on the Kolmogovov complexity theory, information distance (Bennett et al., 1998) is a universal distance metric, which has been successfully applied to many applications. The information distance $D(x, y)$ is defined as the length of a shortest binary program which can compute x given y as well as compute y from x. It has been proved that , up to an additive logarithmic term, $D(x, y) = \max\{K(x|y), K(y|x)\}$. The normalized version of $D(x, y)$, called the *normalized information distance(NID)*, is defined as

$$d_{max}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \quad (1)$$

Parallel to this, the min distance is proposed in (Zhang et al. , 2007), defined as

$$D_{\min}(x, y) = \min\{K(x|y), K(y|x)\}. \quad (2)$$

And the normalized version is

$$d_{min}(x, y) = \frac{\min\{K(x|y), K(y|x)\}}{\min\{K(x), K(y)\}}. \quad (3)$$

3.3 Conditional Information Distance

Conditional information distance is defined as

$$d_{\max}(x, y|c) = \frac{\max\{K(x|y, c), K(y|x, c)\}}{\max\{K(x|c), K(y|c)\}}, \quad (4)$$

$$d_{\min}(x, y|c) = \frac{\min\{K(x|y, c), K(y|x, c)\}}{\min\{K(x|c), K(y|c)\}}. \quad (5)$$

where c is given in both x to y and y to x computation.

The information distance is proved to be universal (Zhang et al. , 2007), that is, if x and y are "close" under any distance measure, they are "close" under the measure of information distance. However, it is not clear yet how to find out such "closeness" in traditional information distance theory. Now the conditional information distance provides a possible solution.Figure 1 gives a more interpretable explanation: the condition c could map the original concepts x and y into different x_c and y_c, thus the variant "closeness" could be reflected by the distance between x_c and y_c, as shown in Figure1.

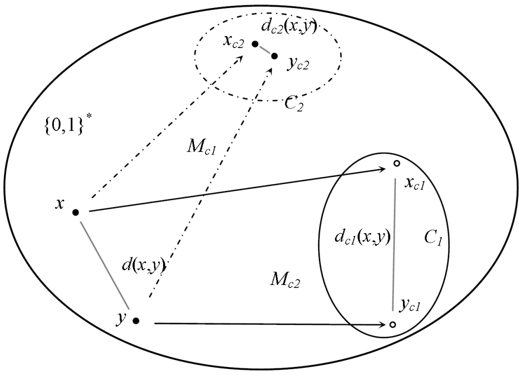

Figure 1: Conditional information distances under different conditions c's

The Kolmogorov complexity is non-computable, that is, to use the information distance measures, we must estimate the K(x) first. There are traditionally two ways to do this: (1) by compression (Li et al. , 2001), and (2) by frequency counting based on coding theorem (Cilibrasi and Vitanyi, 2007). The second approach is implemented in this paper.

4 Answer Validation with Information Distance

Given a question q and a candidate answer c, the answer validation task can be considered as determining the degree of relevance of c with respect to q. The intuition of our approach is that the distance between question and the correct answer is smaller than other candidates. Take the question "What is the capital of the USA?" as an example, among all candidates, the correct answer "Washington" is closest to the question under some distance measure. Thus the answer validation problem is to determine a proper distance measure. Fortunately, it has been proved that the information distance (Bennett et al., 1998) is universal so that the similarity between the question and the answer can surely be discovered using this measure.

Direct calculation of the unconditional distance is difficult and non-flexible. We find it possible and convenient to estimate the conditional information distance between question focus and the answers, under certain context as the condition. As explained previously, different conditions lead to different distance. With the most proper condition and the nearest distance, the best answer can be identified out of previously determined candidates.

The conditional normalized min distance is employed for distance calculation, which is defined

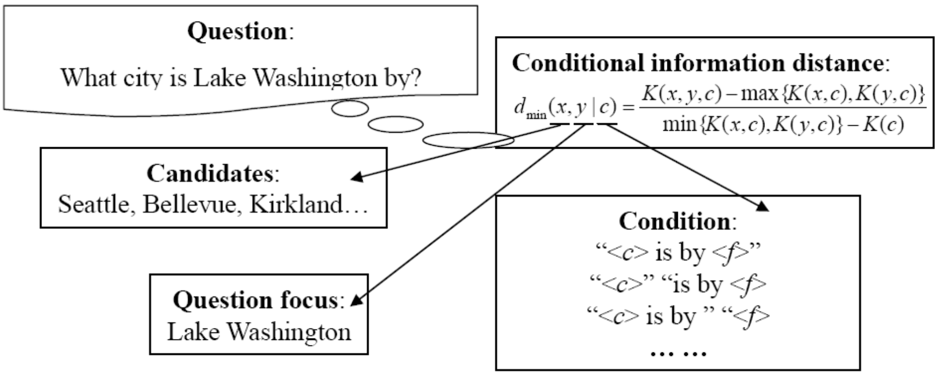
Question:
What city is Lake Washington by?
Conditional information distance:
$d_{\min}(x, y \mid c) = \frac{K(x,y,c) - \max\{K(x,c), K(y,c)\}}{\min\{K(x,c), K(y,c)\} - K(c)}$
Candidates:
Seattle, Bellevue, Kirkland…
Condition:
"<c> is by <f>"
"<c>" "is by <f>
"<c> is by " "<f>
… …
Question focus:
Lake Washington

Figure 2: Sample of conditional information distance calculation.

as:

$$d_{min}(x, y|c) = \frac{K\left(c(x,y)\right) - \max\{K\left(c(x,\phi)\right), K\left(c(\phi,y)\right)\}}{\min\{K\left(c(x,\phi)\right), K\left(c(\phi,y)\right)\} - K\left(c(\phi,\phi)\right)}$$

where x represents the answer candidates, y is the question focus, and c is condition pattern. The function $c(x, y)$ will be described in the Distance Calculation section.

Figure 2 shows the procedure of distance calculation. Given a question and a set of candidates, we calculate the min information distance between question focus and candidates conditioned on surface patterns. Obviously, in order to calculate information distance, there are three issues to be addressed:

1. Question Focus Extraction: since the question answer distance is reformulated as the measure between question focus and answer conditioned on the surface pattern, it is important to extract some words or phrases as question focus.

2. Condition Pattern Generation: Obviously, the generation of the condition is the key part. We have built a well revised algorithm, in which proper conditions can be generated from question sentence according to some heuristic rules.

3. Distance Calculation: after question focus and condition patterns are obtained, the last step is calculating the conditional distance to estimate the relevance between question and answer candidates.

4.1 Question Focus Extraction

Most factoid questions refer to specific objects. A question is asked to learn some knowledge for this object from certain perspective. In our approach, we take the key named entity or noun phrase, usually as the subject or the main object of the question sentence as the reference object. Take the question "What city is Lake Washington by" as example, the specific object is "Lake Washington". The question focus is identified using some heuristic rules as follows:

1. The question is processed by shallow parsing. All the noun phrases(NP) are extracted as NP set.

2. All the named entities(NE) in the question are extracted as NE set.

3. If only one same element is identified in both NE and NP set, this element is considered as question focus.

4. If step 3 fails, but two elements from NE and NP set have overlap words, then choose the element with more words as question focus.

5. If step 3 and 4 fail, choose the candidate, which is nearest with verb phrase in dependency tree, as question focus.

4.2 Condition Pattern Generation

A set of hierarchical patterns is automatically constructed for conditional min distance calculation.

4.2.1 Condition Pattern Construction

Several operations are defined for patterns construction from the original question sentence. We describe pattern set construction with a sample question "What year was President Kennedy killed?":

1. With linguistic analysis, the question is split into pieces of tokens. These tokens in-

clude wh-word phrases, preposition phrases, noun phrases, verb phrases, key verb, etc. The example question is split into “What year”(wh-word phrase), “was”(key verb) “President Kennedy” (noun phrases), “killed”(verb phrase).

2. Replace the wh-word phrases with the candidate placeholder $\langle c \rangle$. Then the words “What year” is replaced with placeholder $\langle c \rangle$.

3. Replace the question focus with the focus placeholder $\langle f \rangle$, and add this pattern to the pattern set. The example question focus is identified as “President Kennedy”. It is replaced with placeholder $\langle f \rangle$. The first pattern “$\langle c \rangle$ was $\langle f \rangle$ killed?” is generated.

4. Voice Transformation: with morphology techniques, verbs are expanded with all their tense forms (i.e. present, past tense and past participle). The tokens’ order is adjusted to transform between active voice and passive voice. Both patterns are added to the patterns set. For sample question, the passive pattern is translated into active pattern, “$\langle c \rangle$ kill $\langle f \rangle$”.

5. Preposition addition: for time and location questions, the preposition (i.e. in, on and at) is added before the candidate $\langle c \rangle$; Then the pattern “$\langle c \rangle$ was $\langle f \rangle$ killed” is reformulated as “(in |on) $\langle c \rangle$ was $\langle f \rangle$ killed”.

6. Tokens shift: preposition phrase token could be shifted to the begin or the end of pattern, and “key verb” must be shift before the “verb phrase”. Then the pattern “(in |on) $\langle c \rangle$ was $\langle f \rangle$ killed” can be reformulated as “$\langle f \rangle$ was killed (in |on) $\langle c \rangle$”.

7. Definitional patterns: several heuristic patterns, as introduced at (Hildebrandt et al. , 2004), are added into our final pattern sets, such as “$\langle c \rangle$, $\langle f \rangle$”.

By such heuristic rules, the original pattern set is obtained from question sentence. The patterns are initially enclosed in quotation marks, which means exact matching. However, by eliminating these quotations, or reducing the scope that they cover, the matching is relaxed as words co-occurrence. The patterns are expanded into different strict-level patterns by adding or removing quotation marks for each tokens or adjacent tokens combination. Several condition pattern samples are shown in Table 1

Table 1: Sample condition patterns, ‘ “” ’ denotes exact match in web query.

① “ <f>(was | were) killed (in | on) <c>”
② “ (in | on) <c>, <f>(was | were) killed”
③ “ (in | on) <c>” **&** “<f>(was | were) killed”
④ “ (in | on) <c>” **&** “<f>” **&** “(was | were) killed”
⑤ in | on <c><f>(was | were) killed

Each operation introduced above is given a predefined confidence coefficient(cc). Then the confidence coefficient of a pattern is defined as the multiplication of cc for all performed operations to generate this pattern.

4.2.2 Condition Pattern Ranking

From the previous step, a set of condition patterns and corresponding confidence coefficient are obtained. Let p_i denotes the ith pattern in the pattern set, and cc_i is the confidence coefficient for the ith pattern. The confidence coefficient estimation in previous section contains much noise. And the patterns with similar confidence coefficient make little difference. Therefore, the exact confidence coefficient value is not directly used. We cluster the patterns into different priority groups. C_j denotes the pattern cluster with jth priority. Here, the smaller j means higher priority. The condition patterns are ranked mainly based on confidence coefficient and the number of double quotation marks. The following algorithm shows each step in detail:

Table 2: patterns ranking algorithm

Input	patterns set $C = \{(p_i, cc_i)\}$
Algorithm	
(1)	Initialize $C_j = \emptyset, j = 0$
(2)	if C is empty, end this algorithm
(3)	Select (p_{max}, cc_{max}), where $cc_{max} \geq cc_i, (p_i, cc_i) \in C$
(4)	if C_j is empty, add cc_{max} into C_j, jump to (2)
(5)	select the minimum confidence coefficient (p_{min}, cc_{min}) from C_j, compare it with (p_{max}, cc_{max}). if the number of double quotes(“”) in p_{min} is equal to the number in p_{max}, add p_{max} into C_j. otherwise, $j = j + 1, C_j = \{p_{max}\}$.
(6)	jump to (2) and repeat

4.3 Distance Calculation

Conditional min distance d_{min} is used to measure the relevance between question and candidate. From section 3, d_{min} is not computable, but approximated by frequency counts based on the coding theory:

$$
\begin{aligned}
&d_{min}(x,y|c)\\
&=\frac{K\Big(c(x,y)\Big)-\max\{K\Big(c(x,\phi)\Big),K\Big(c(\phi,y)\Big)\}}{\min\{K\Big(c(x,\phi)\Big),K\Big(c(\phi,y)\Big)\}-K\Big(c(\phi,\phi)\Big)}\\
&=\frac{\log f\Big(c(x,y)\Big)-\min\{\log f\Big(c(x,\phi)\Big),\log f\Big(c(\phi,y)\Big)\}}{\max\{\log f\Big(c(x,\phi)\Big),\log f\Big(c(\phi,y)\Big)\}-\log f\Big(c(\phi,\phi)\Big)}
\end{aligned}
$$

The function $c(x, \emptyset)$ means substituting $\langle c \rangle$ in c by answer candidate x and removing placeholder $\langle f \rangle$ if any. Similar definition applies to $c(y, \emptyset)$, $c(x, y)$. For example, given pattern "$\langle f \rangle$ was invented in $\langle c \rangle$", question focus "the telegraph" and a candidate "1867". $c(x, \emptyset)$ is "was invented in 1867". $c(y, \emptyset)$ is "the telegraph was invented", and $c(x, y)$ is "the telegraph was invented in 1867". The frequency counts $f(x)$ are estimated as the number of returned pages by certain search engine with respect to x . $f(c(\phi, \phi))$ denote the total pages indexed in search engine. Two types of search engines "Google" and "Altavista" are employed.

The patterns are selected in priority order to calculate the information distance for each candidate.

5 Experiment and Discussion

5.1 Experiment Setup

Data set: The standard QA test collection (Lin and Katz, 2006) is employed in our experiments. It consists of 109 factoid questions, covering several domains including history, geography, physics, biology, economics, fashion knowledge, and etc.. 20 candidates are prepared for each questions. All answer candidates are first extracted by the implemented question answering system. Then we review the candidate set for each question. If the correct answer is not in this set, it is manually added into the set.

Performance Metric: The top 1 answer precision and mean reciprocal rank (MRR) are used for performance evaluation.The top 1 answer means the correct answer ranks first with our distance calculation method, and $MRR = \frac{1}{n} * \sum_i (\frac{1}{rank_i})$, in which the $\frac{1}{rank_i}$ is 1 if the correct answer occurs in the first position; 0.5 if it firstly occurs in the second position; 0.33 for the third, 0.25 for the fourth, 0.2 for the fifth and 0 if none of the first five answers is correct.

The open source factoid QA system ARANEA (downloaded from Jimmy Lin's website in 2005) is used for comparison, which implements an approximate $tfidf$ algorithm for candidate scoring. Both ARANEA and our proposed approaches use the internet directly. Google is used as the search engine for ARENEA, and our conditional normalized min distance is calculated with Google and Altavista respectively.

5.2 Experiment Results

The performances of our proposed approach and ARANEA are shown in Table 3. For top 1 answer precision, our conditional min distance calculation method through Google achieves 69.7%, and Altavista is 66.1%, which make 56.6% (69.7% v.s.42.2%) and 50.0% (66.1% v.s 42.2%) improvement compared with ARENEA's $tfidf$ method. Our proposed methods achieve 0.756 and 0.772 compared with ARENEA's 0.581 for MRR measure.

Table 3: Performance comparison, where $d_{min}(G)$ denotes the distance calculation through "Google", $d_{min}(A)$ through "Altavista"

	tfidf	$d_{min}(G)$	$d_{min}(A)$
# of Top 1	46	72	69
% of Top 1	42.2	69.7	66.1
MRR	0.581	0.772	0.756

Table 4 shows some correct answer validation examples. the Google Condition(GC) and the Altavista Condition(AC) columns are the employed condition patterns for distance calculation. For question 1400, the conditional normalized google min distance calculates the distance between question focus "the telegragh" and all 20 answer candidates. The minimum distance score is achieved between "the telegraph" and "1837" with the condition pattern "$\langle f \rangle$ was invented in $\langle c \rangle$". Therefore, the candidate "1837" is validated as the correct answer. Meanwhile, the minimum value for conditional normalized altavista min distance is achieved on the same condition.

These results demonstrate that the distance calculation method provides a feasible solution for answer validation.

In discussion section, we will study three questions:

1. What is the role of search engine?
2. What is the role of condition pattern?
3. What is the role of question focus?

Table 4: Question Examples in conditional information calculation through Google and Altavista. GC:Google Condition; AC:Altavista Condition

ID	Question	GC	AC	Answer	Question focus
1400	When was the telegraph invented?	"?y was invented in ?s"	"?y was invented in ?x"	1837	the telegraph
1401	What is the democratic party symbol?	"?y is ?x"	"?y is ?x"	the donkey	the democratic party symbol
1411	What Spanish explorer discovered the Mississippi River?	"?x discovered ?y"	"?x" "discovered" "?y"	Hernando de Soto	the Mississippi River
1412	Who is the governor of Colorado?	"?y is ?x"	"?y, ?x"	Gov. Bill Ritter	the governor of Colorado
1484	What college did Allen Iverson attend?	"?y attended ?x"	"?x" "did ?y"	Georgetown University	Allen Iverson attend

5.3 Discussions

5.3.1 Role of Search Engine

The rise of world-wide-web has enticed millions of users to create billions of web pages. The redundancy of web information is an important resource for question answering. Our Kolmogorov Complexity based information distance is approximated with query frequency obtained by search engine. Two types of search engines "Google" and "Altavista" are employed in this paper. The number of top 1 correct answer is 72 through "Google" and 69 through "Altavista". There is little difference between two numbers, which shows that the information distance based on Kolmogorov Complexity is independent of special search engine. The performance didn't vary much with the change of search engine. Actually, if the local data is accumulated large enough, the information distance can be approximated without the internet. The quality and size of data set affect the experiment performance.

5.3.2 Role of Condition Pattern

Pattern set offers convenient and flexible condition for information distance calculation. In the experiment, there are 61 questions correctly answered by both Google and Altavista. 46 questions of them employ different patterns. Considering Question 1412, the condition pattern in Google is "$\langle c\rangle$ is $\langle f\rangle$", while in Altavista, it is "$\langle f\rangle$, $\langle c\rangle$". However, the correct answer "Gov. Bill Ritter" is identified by both methods. The information distance is stable over specific condition patterns.

5.3.3 Role of Question Focus

Question focus is considered as the discriminator for the question. The distance between a question and a candidate is reformulated as the distance between question focus and candidate conditioned on a set of surface patterns. The proposed approach may not properly extract the question focus, but the answers can be correctly identified when the condition pattern becomes loose enough. Take the question 1484 "What college did Allen Iverson attend?" as example, the verb "attend" is tagged as "noun", then question focus is mistakenly extracted as "Allen Iverson attend", instead of the correct "Allen Iverson". The two conditional information distance method still identify the correct answer "Georgetown University". Because they both employed the looser condition patterns '"$\langle c\rangle$" "$\langle f\rangle$"' and '"$\langle c\rangle$" did "$\langle f\rangle$"'.Therefore, our proposed distance answer validation methods are robust to the question focus selection component.

From the discussion above, it can be seen that our algorithm is stable and robust, not depending on the specific search engine, condition pattern, and question focus.

6 Conclusions

We have presented a novel approach for answer validation based on information distance. The answer validation task is reformulated as distance calculation between question focus and candidate conditioned on a set of surface patterns. The experiments show that our proposed answer validation method makes a great improvement compared with ARANEA's *tfidf* method. Furthermore, The experiments show that our approach is stable and robust, not depending on the specific search engine, condition pattern, and question focus. In future work, we will try to calculate information distance in the local constructed data set, and expand this distance measure into other application fields.

Acknowledgement

This work is supported by National Natural Science Foundation of China (60572084, 60621062), Hi-tech Research and Development Program of China (2006AA02Z321), National Basic Research Program of China (2007CB311003).

References

Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi, and Richard J.Mammone. 2001. *Question answering using maximum entropy conmponents*. In Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics on Language tecnologies.

Anselmo Penas, A. Rodrigo, F. Verdejo. 2007. *Overview of the Answer Validation Exercise 2007*. Working Notes for the CLEF 2007 Workshop.

C.H. Bennett, P. Gacs, M. Li, P. Vitányi, W. Zurek.. 1998. *Information Distance*. *IEEE Trans. Inform. Theory*, 44:4, 1407–1423.

Eric Brill and Susan Dumais and Michele Banko. 2002. *An analysis of the AskMSR question-answering system*. EMNLP '02: the ACL-02 conference on Empirical methods in natural language processing.

Hildebrandt W., Katz B., and Lin J. 2004. *Answering Definition Questions Using Multiple Knowledge Sources*. Proceedings of Human Language Technology Conference. Boston, USA.

Jeongwoo Ko, Luo Si, Eric Nyberg. 2007. *A Probabilistic Graphical Model for Joint Answer Ranking in Question Answering*. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.

Jimmy Lin and Boris Katz. 2006. *Building a reusable test collection for question answering*. J. Am. Soc. Inf. Sci. Technol..

Jimmy Lin. 2007. *An Exploration of the Principles Underlying Redundancy-Based Factoid Question Answering*. ACM Transactions on Information Systems, 27(2):1-55.

Jinxi Xu, Ana Licuanan and Ralph Weischedel. 2003. *Trec 2003 qa at bbn: Answering definitional questions*. In Proceedings of the 12th Text REtrieval Conference, Gaithersburgh, MD, USA.

Ming Li and Paul MB Vitanyi. 1997. *An Introduction to Kolmogorov Complexity and Its Applications*. Working Notes for the CLEF 2007 Workshop.

M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang.. 2001. *An information-based sequence distance and its application to whole mitochondrial genome phylogeny*. Bioinformatics, 17:2.

M. Subbotin and S. Subbotin. 2001. *Patterns of Potential Answer Expressions as Clues to the Right Answers*. In TREC-10 Notebook papers. Gaithesburg, MD.

R. Cilibrasi, P.M.B. Vitányi. 2007. *An Exploration of the Principles Underlying Redundancy-Based Factoid Question Answering*. EEE Trans. Knowledge and Data Engineering, 19:3, 370–383.

Shen, Dan and Dietrich Klakow . 2006. *Exploring correlation of dependency relation paths for answer extraction*. In Proceedings of COLING-ACL, Sydney, Australia.

Xian Zhang, Yu Hao, Xiaoyan Zhu, and Ming Li. 2007. *Information Distance from a Question to an Answer*. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.