IJARESM
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING, SCIENCE & MANAGEMENT

EVALUATION OF COMPOSITIONAL DISTRIBUTIONAL SEMANTIC MODEL ON QUESTION ANSWERING SYSTEM WITH MULTIPLICATION OPERATOR

Rutal Mahajan[1]
Assistant professor, Computer engineering department, SNPIT & RC, Umrakh, Bardoli, Gujarat, India[1]

***Abstract*: A question answering (QA) system provides direct answers to user questions by consulting its knowledge base. In recent decade, there is an explosion in the freely available information, improvements in information technology and the increase in people's desire for the better information access. One question arises related to information "meaning" while giving the exact answer of the question instead of the large document freely available information, which has created the emergence of a wide variety of computational models for modeling the meaning of words based on the underlying assumption of the distributional hypothesis, that the meaning of word can be inferred from its use. This paper explores the role of distributional semantic models (DSM) in Question Answering (QA) system. The idea is to compute the relatedness between the user question and the candidate answer retrieved by the search module. One of the shortcomings of DSMs found while applying for larger linguistic unit is "bag-of-words" limitation. A result of semantic compositionality is boost up when used with multiplication operator.**

***Index Terms*— Distributional Semantic Model, Question Answering (QA) system, semantic compositionality**

INTRODUCTION

Question answering is a fast growing research area which has been developed past few years in many open evaluation campaigns. The aim of the Question Answering system is to examine user question and to provide most appropriate answer to that question, instead of providing collection of documents which may contain answer of user question. Based on the relevance between the user question and the document or passage containing answer, the candidate answers are retrieved from which most appropriate answer to the user question is extracted. Due to the increased desire of people to get better information access, in [1] various standards and issues are discussed, which are necessary to be considered in the design of the QA system.

Distributional Semantic Models (DSMs) represents the word meaning through linguistic contexts. Thus it could be useful to deal with the one of the issue related to QA system design, namely, query language complexity. It relates to the difficulty of extracting the intended meaning from linguistic expression. In terms of QA system, it can be represented as

the problem of processing the information request in various ways. e.g. the question "Who is the national bird of India?" same can be asked in assertive manner "tell me the name of the national bird of India". Such problem of richness of language can be easily handled by DSMs thus it is evaluated here on QA system. DSMs works on the base of the distributional hypothesis that, "the meaning of word can be inferred from its use" [2]. Using this hypothesis the meaning of the word can be expressed by geometrical representation in semantic space. In semantic space a word is represented using a vector. A word vector is built by analysing the contexts in which it occurs in the reference text. The contexts can be a sentence, document, or window of surrounding terms. Different parameters required to build the DSMs are given in [3]. The DSMs differs primarily with regards to these parameters, such as, context type, context window size, dimensionality reduction technique, similarity measure and frequency weighting scheme.

The significant advantage of such methods of distributional semantics is that the meaning representations can be computed automatically from given large volume of text and also they are independent of corpus language so here we have applied it to the open domain QA system. It has many practical uses in various linguistic and cognitive tasks such as synonym test [4], automatic thesauri construction [2], word sense disambiguation [5] and bilingual information extraction [6]. This paper is intended to use DSM for QA system, which has not been so famous yet. In QA system, using the paradigmatic relations between words, the similarity between candidate answers and user question is calculated.

To test the effectiveness of the distributional semantic model for QA system, we have evaluated on the existing open domain QA system OpenEphyra[7]. OpenEphyra is a QA system developed by the one of the developer of IBM Watson [8]. OpenEphyra is the integration between several techniques for question analysis, query generation and answer extraction. Answer extraction is done on the base of scores assigned to the candidate answers by pipeline of filters. Individual approaches often suffer from low precision. Thus the system's design was based on using the advantages of various individual approaches to build a big system and increase the overall performance.The highly modular design of openEphyra makes it suitable for this task.

The idea to use DSM in QA system is to build the distributional filter to the answer selection pipeline of openEphyra[7] [4]. Here mainly evaluation is carried out on spaces: term-term co-occurrence matrix (TTM), Latent Semantic Analysis (LSA) applied on TTM, Random indexing (RI) to reduce dimension of TTM, Hyper space Analogous to Language (HAL).

DISTRIBUTIONAL SEMANTIC MODELS

We have constructed the DSM on the co-occurrence matrix using sliding window w of co-occurring words. Here we have used the list of candidate answers retrieved by the search engine to build the reference corpus and the vocabulary V, to build the n x n co-occurrence matrix M, whose coefficients m_{ij} are number of co-occurrence counts of the words within window size.

The term-term matrix M is based on simple word co-occurrences and it represents the simplest semantic space known as term-term co-occurrence matrix. Here many methods are used to perform dimensionality reduction and to derive higher order relations. Among those, we have used LSA, RI, HAL.

Latent Semantic Model

Latent Semantic Analysis (LSA) [9] is one of the most-well known vector space model. LSA uses Singular Value Decomposition (SVD) to find a reduced vector space that fits the original as well as a possible lower ranked matrix. SVD decomposes the original co-occurrence matrix M of dimensions T x N, with T being the total number of terms and N the total number of documents, into the product of three matrices $U \sum V^T$, i.e. $M= U \sum V^T$.

Where $U^T U = V^T V =1$ (i.e. U and V are orthonormal matrices whose columns are right and left eigen vectors of the matrices M^TM and MM^T respectively) and $\sum$ is a diagonal matrix of singular values. Dimension of U is T x m, dimension of $\sum$ is m x m and of V is m x N.

The dimension reduction can also be performed using the kth approximation of M by choosing the top k singular values. While considering k- approximation of M, the complexity of computing similarity between terms is reduced.

Hyperspace Analogous to Language

In 1996, Lund and Burgess described a framework [10] in which the occurrence of words in a window of 10 words surrounding each target word are counted and weighted in a way which is inversely proportional to its distance from the target word. By moving the window over the corpus in one word increments, and counting co-occurrence statistics, a co-occurrence matrix can be formed. The matrix actually records words that appear before and after the target word and in doing so it introduces some degree of order information. Various types of dimensional reduction or feature selection may then be performed and various measures of similarity used.

Random Indexing

Random Indexing [11] (RI) is introduced as an effective and scalable method for constructing DSMs from large volumes of text. It is based on the concept of Random Projection; according to it the chosen high dimensional vectors are nearly orthogonal.

Random projection can be defined for the given matrix M of n x m dimension and m x k matrix of R, the new matrix M' of n x k dimensions is

$$M_{txd}\ R_{dxk} = M'_{txk} \quad k<<m$$

The method is scalable because it performs a type of implicit dimensional reduction and it performs this in an incremental fashion. Random indexing works as follow:

1. For a given corpus a random index vector is assigned to each term. Random index vector is high dimensional sparse vector of ternary values (-1,0,1)
2. Context vector of the term is given by summing up the index vector of co-occurring terms within predetermined context.

INTEGRATION OF DSMS INTO QA SYSTEM

Distributional Semantic Models (DSMs) represent The DSMs are integrated into answer selection module into existing QA system OpenEphyra [7]. Here it will work as the distributional filter for answer re-ranking. To get the DSMs work properly with QA system, the inputs to it should be pre-processed and compositionality approach should also be added to work with QA system. Architecture of the complete QA system OpenEphyra is highly modular so change is at only answer selection module after integration. Due to modular design of the answer selection module, the distributional semantic model can also be easily added or removed as a distributional filter as and when needed, which aims at computing similarity between user's question and each candidate answer retrieved by search module.

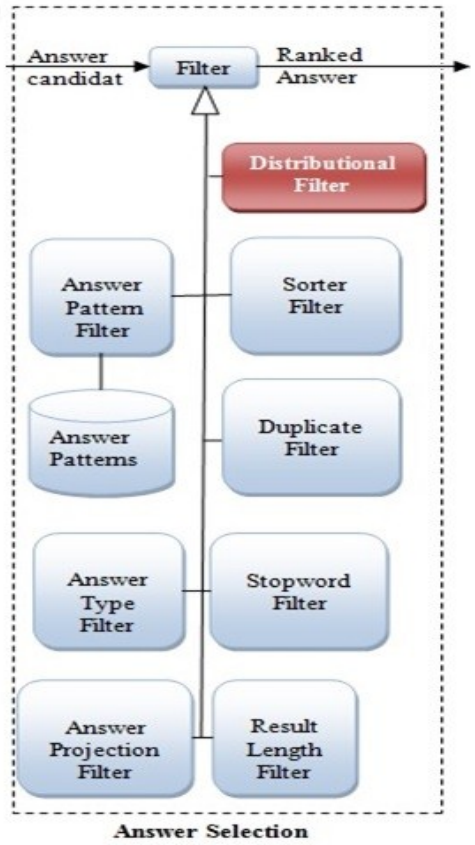

Fig. 1. Integration of distributional filter in answer selection module of OpenEphyra

First the question string is received at the question interpretation phase, where the question string is normalized, e.g. we drop unnecessary tokens and stem verbs and nouns. The Question Analyzer then determines the expected answer type of the question and interprets the question to create a more concise representation of the question string. After this a set of query generators transforms the question into queries for document retrieval. The search results are passed through a set of filters to create a ranked list of answers. The distributional filter is added at answer selection module measures the similarity between all the candidate answers from search module and a question according to process of the particular distributional semantic model selected. Candidate answers are ranked according to their similarity scored given by distributional filter.

Different Distributional filters adapted in this QA system are TTM, LSA, HAL and RI. To use these filters, at a time any single filter among these can be added in to the answer selection module to be used for candidate answer re-ranking.

Commonly in all DSMs, the words are represented by vectors such as $v = (v_1\ v_2 \ldots\ v_n)^T$ and $u = (u_1\ u_2 \ldots\ u_n)^T$. so the similarity between words are calculated by cosine angle between them.

$$Cos(u,v) = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2 \sum_{i=1}^{n} v_i^2}} \qquad (1)$$

Now to integrate DSM in Question answering system for re-ranking of candidate answers according to their similarity with question sentence, a technique to combine the word vector of candidate answers words and question sentence words is needed. The simplest approach, a point wise addition operator is used in [4] to compose sentence vector by combining the word

vectors. Point wise sum of components given in [4] is defined as:

$$p_i = u_i + v_i \quad (2)$$

if question $q= q_1q_2...q_n$ and candidate answer $a_i = a_1a_2 \ ... \ a_m$ are respective sentences, their vector representations in the semantic space can be build using addition operator on vector representation of words belonging to them:

$$q = q_1+ q_2+q_3+...+q_n$$

$$a = a_1+ a_2+ a_3+...+a_n \quad (3)$$

Similarity between q and a is calculated by cosine similarity measure same as in eq.(1). Similarity score is added to the candidate answers.

Same way here in this work, we have evaluated the OpenEphyra QA system for answer re-ranking task using point-wise multiplication operator, where point-wise multiplication is defined as:

$$q = q_1 . q_2 . q_3. \ ... \ . q_n$$

$$a = a_1 . a_2 . a_3 . \ ... \ .a_n \quad (4)$$

Though, point-wise multiplication operator is also commutative as addition operator but it only adds information which is present in both the vectors. Thus by using multiplication operators the results are boost up with higher co-occurrence information.

I. EVALUATION AND RESULTS

In this work, the developed Distributional semantic models with compositionality are evaluated on the Open Domain Question Answering System. As the primary goal is to check the effectiveness of the Distributional semantic model in Question Answering task, here the evaluation is conducted on the existing question answering system, OpenEphyra, which has given state-of-the art performance in TREC evaluation campaigns.

In this evaluation task, the different types of questions are asked to the QA system and it has to return the most appropriate answer for that question.

Data Set

Here distributional semantic model which is added to existing question answering system has to map the user question and the possible candidate answers given by the previous module of the QA system on to the vector space, and has to find the similarity between the question and candidate answer along with assigning similarity scores to them. Based on the similarity scores assigned to the candidate answers they should be re-ranked. To check the effectiveness of the DSMs in QA system the performances of existing baseline (OpenEphyra) system and the system with DSM is compared here. Effect of different compositionality operator is also compared here. As the evaluation is carried out over web based open domain question answering system, the candidate answers are retrieved from the web using Google custom search module, from which here we created text corpus. The evaluation is performed on set of 50 web based questions, which are from TREC QA data and from general knowledge question quiz book. The question set contains interrogative question of who type, how type questions and also non-interrogative type questions. Also In this evaluation, some parameters of Distributional semantic models need to be set. Sliding window of size 4 is considered as a context to count co-occurrence. Number of reduced dimensions used in LSA and RI is 300.

Experimental Setup

The experimental setup includes various external libraries, setting of different model parameters and the configuration of Google custom Search engine for proper working of baseline system OpenEphyra on web based data.

Results

Comparisons of different DSMs, when they are used with multiplication operators, are described in figure 2 to figure 3. All DSMs performs much better when they are used with multiplication operator, except for the case of RI, alone as well as combined with baseline. This drastic change of result is because of the multiplication operator. It scales the non-zero values of scores for result very well so that we get more answers re-ranked at top-1 or top-5 positions.

In the case of RI multiplication operator ignores the non-zero result and gives the result zero if any of the two vectors contain zero. Figure 2 clearly describes this case. But RI +baseline has given answers even with multiplication operator, it is because of combined effect of baseline filters.

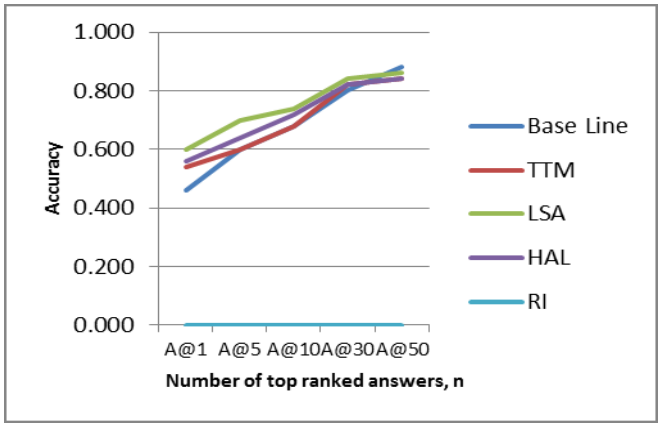

Fig. 2. Accuracy comparison of DSMs and a Baseline system @ n with multiplication compositionality operator

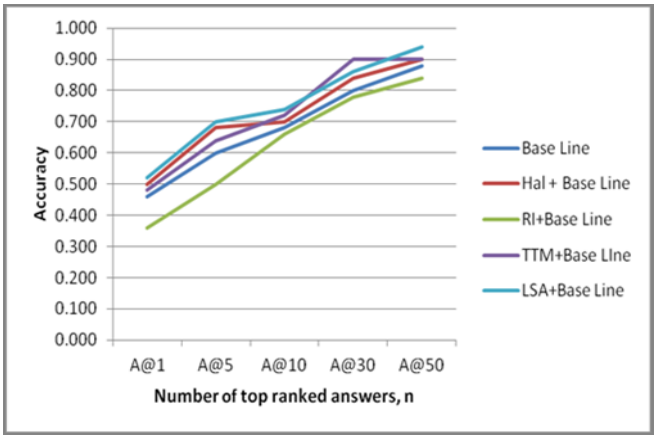

Fig. 3. Accuracy comparison of integrated DSMs and a Baseline system @ n with multiplication compositionality operator

CONCLUSION

Due to increased need of Question answering system in real world applications, it is necessary to build the QA system which not only give the answer as fast as possible but also gives the correct answer. In this research different distributional semantic models and compositionality operators on QA system are evaluated and successfully used them in web based QA systems for re-ranking the answers. The results of different distributional semantic models (combined with baseline and alone) proved their effectiveness in this task. Among all the DSMs we evaluated, we found that the model Random Indexing is not appropriate model with multiplication operator but if it is used with baseline system then it gives significant results.

LSA with multiplication operator is more sufficient among the all model when it is used as alone system. But TTM + Baseline system with multiplication operator is more appropriate system with multiplication operator to be used for real world application of QA.

REFERENCES

[01] David Ferrucci, Eric Nyberg, James Allan, Ken Barker, Eric Brown, Jennifer Chu-Carroll, Arthur Ciccolo, Pablo Duboue, James Fan, David Gondek, Eduard Hovy, Boris Katz, Adam Lally, Michael McCord, Paul Morarescu, Bill Murdock, Bruce Porter, John Prager, Tomek Strzalkowski, Chris Welty, Wlodek Zadrozny, "Towards the Open Advancement of Question Answering Systems", IBM research report, In RC24789, pp. 1-29, URL: http://research.ijcaonline.org/volume65/number12/pxc3886122.pdf , 22 April, 2009, [Accessed: Dec. 09, 2013]

[02] Lance De Vine, "Some extensions to representation and encoding of structure in models of distributional semantics", Master's Thesis, Queensland University, Australia, June 2013.

[03] D. Kiela and S. Clark, "A Systematic Study of Semantic Vector Space Model Parameters," In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality, Gothenburg, Sweden, pp. 21-30.,2014.

[04] Piero Molino, Pierpaolo Basile, Annalina Caputo, Pasquale Lops, Giovanni Semeraro, "Exploiting Distributional Semantic Models in Question Answering", IEEE Sixth International Conference on Semantic Computing, pp. 146-153,19-21 September, 2012

[05] H. Schütze, "Automatic word sense discrimination." Computational linguistics-Special issue on word sense disambiguation, vol. 24, no. 1, pp. 97–123.1998.

[06] D. Widdows, & S. Cederberg, "Monolingual and bilingual concept visualization from corpora." In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4, pp. 31–32.2003

[07] Nico Schlafer. "Deploying semantic resources for open domain question answering." Master's thesis, Universitat Karlsruhe, May 2001.

[08] D. Ferrucci, E. Brown, C. Jennifer, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. Murdock, E. Nyberg, J. Prager, N.Schlaefer, C. Welty, "Building Watson: an overview of DeepQA Project.", AI Magazine, Vol. 31, no. 3.,2010.

[09] T. Landauer, P. Foltz, & D. Laham, "Introduction to Latent Semantic Analysis." Discourse Processes, 25, pp. 259-284.1998.

[10] K. Lund, and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence." Behavior Research Methods, Instruments, & Computers, vol. 28, no. 2, pp. 203–208. 1996.

[11] P. Kanerva, J. Kristofersson, and A. Holst, "Random indexing of text samples for latent semantic analysis." In Proceedings of the 22nd Annual Conference of the Cognitive Science Society Vol. 1036, 2000.

