A proposed algorithm for identifying words relations depending on co-occurrences

Heba Saied Abd Al-Rahman Konswah
Ghada A. ElKhayat
Safaa Hussien
Information Systems and Computers Dept. Faculty of Commerce,
Alexandria University, Alexandria Egypt

Mahmoud Youssef Ahmed
*Management Science & Information Systems Dept. Rutgers Business School,
Rutgers University, New Jersey, USA

Keywords
Co-occurrences, Information Extraction, Information Retrieval, Ontology Construction, Semantic Networks, Text Processing.

Abstract

Information extraction needs text analysis prior to processing to identify main words and the frequently repeating words that combines them. This process is often done using mining tools with manual verification or fully manual, in this paper we propose an algorithm based on co-occurrences frequencies for identifying the most frequent important words in a specific domain and the technique for building their relations to create an informative words network that can be used in ontology construction or information retrieval processes.

Introduction

Text processing has wide range of applications, the common process between them is the pre-processing analysis that aims at knowing the main expressions used in the selected domain. This is a core prior step that produces the input for the following stages, as it helps in recognizing the text features and structure.

Each domain has its own terms that are frequently used and has a main impact on forming the information about this domain. These terms comes in different sentences in many structures, it happens to have more than word that combines it in most occurrences. This shows that those words are related, and this relation can be represented numerically by the frequencies of their occurrences. The comparison between a main word occurrences frequency and another frequent word that often combines it could give an indication about a significant relation. This is a proposed technique to use for exploring the text corpus; it could be an aiding factor to be considered with other statistical approaches or to be validated by human experts.

There are many applications that are based on similar input as starting point including: Ontology construction and updating, Information Extraction, Search, Classification, Summarization and Knowledge Base Building. All these applications are based on a structure that represents the text features and relations, in a form that could be a semantic network, keywords or rules which is considered our final targeted output from this proposed technique.

The rest of this paper is organized as follows, the next section is presenting the related work, and section two contains detailed explanation for the proposed algorithm components. The third section shows the conclusion and future work.

Related Work:

Al-Khalil ontology (Aliane et al., 2010) used statistical methods namely the repeated segments calculations. The problem in depending on repeated segments method is it needs to find an exact text match to count on, which doesn't represent the real life situation in. As in most cases we may find a sentence containing similar words and deliver the same meaning but each time it comes in different

structure even with light changes.

Alanazi et al. (Alanazi et al., 2015) uses frequency analysis to find informative words in the medical domain and collocations, which is close to co-occurrences method but it only refers to two sequenced words that are often used together. This research uses collocations to find entities related to diseases names and treatments.

Mazari et al. (2012) applied two statistical methods, repeated segments and co-occurrences. Repeated segments approach is used to identify candidate terns while constructing the ontology, in this research the segment is set to 4 words and a threshold is randomly selected depending on the corpus size (100 for simple word and 20 for compound term). Co-occurrences approach is used to build the ontology relations and to update the ontology, but in this research the co-occurrences is measuring only pairs of words.

A similar research uses co-occurrences statistical method to build contextual relations is (Bounhas et al., 2011). It uses weighted approach that links terms to their corresponding roots and sub-related nodes in cases of specific domain ambiguous terms. Chang et al. (2007) used heuristic rules to extract knowledge elements candidates which represents sentence that has meaningful terms. A set of 500 learning resource in the computer science domain were collected from internet and manually classified. Extracting knowledge elements candidates and terms using heuristic rules, with setting a fixed length for a sentence which is not recommended as it could lead to inaccurate results. Classification methods including Decision trees, Support Vector Machine and Naïve Bayesian were used to classify the knowledge elements to their semantic type category. Researches were conducted for similar purpose but they have applied different approaches, such as statistical approaches; (Benajiba et al., 2009) introduced an approach that explores contextual and lexical features using three discriminative machine learning frameworks: Support Vector Machine (SVM), Maximum Entropy (ME), and Conditional Random Fields (CRF).

Some researchers depended on mapping words relations into another translated language existing relations; (Abdelali et al., 2003) used English lexicon ontology to translate English queries to Chinese and Arabic concepts also (Zaidi et al., 2005) used Arabic words mapped to English words relations for query expansion purpose in the legal domain. Abouenour et al. (2010) investigates the Arabic Word Net (AWN) Named Entities (NEs) enrichment by using YAGO ontology. Also (Alkhalifa et al., 2009) introduced an approach for extending AWN Geographical NEs coverage using Arabic Wikipedia (AWP). Jarrar (2011) used the mapped relations for constructing Arabic ontology semantic relations, by mapping the mined Arabic concepts to AWN inherited relations.

This work can be classified as ontology construction research according to (Konswah, 2016) model see Fig.1.

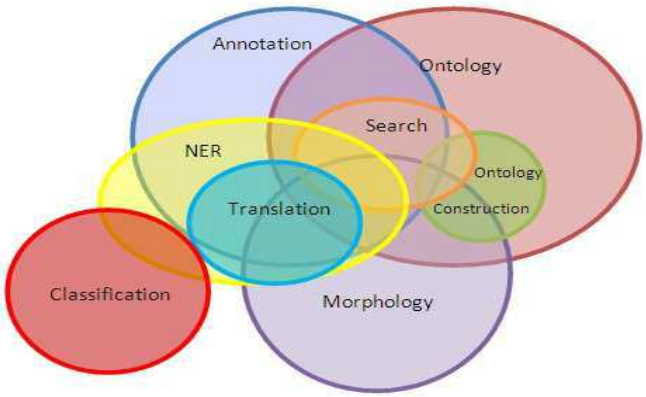

Fig. 1 Modeling Arabic Language Research Intersections (Konswah, 2016)

Proposed Algorithm

The proposed algorithm presents a technique that is based on calculating co-occurrences

between frequent words. This process is done through many steps, starting with preprocessing segmentation followed by calculating words frequencies with clarification example.

Segmentation

The processing corpus should be first tokenized and segmented. *Tokenization* is to separate each text word, symbol or number into tokens. *Segmentation*: to get the related words frequencies we need to identify the context boundaries, which is considered a sentence/segment. A sentence could indicate that the inner words are related; as if a word comes in a sequence after another word but in a different sentence, they might not be related at all. A sentence boundary could be defined by dot, comma or a number of words. The dot stops the sentence and the comma could start a new one, the number of words is to put a limit of words number to each sentence, which is not preferably used as it won't give accurate results but we assume it is a solution in cases of un-punctuated text corpus.

Calculating Frequencies

In this step we may use one of two alternatives: a) feeding the system with domain important words as input to calculate their frequencies and the words related to these input words frequencies, b) calculating all tokens frequencies and check the most frequent words relations.

A) Feeding main words to the system:

Suppose we know a list of core words that we are interested to get the words related to them. These words can be pre-defined by a domain expert and after text corpus analysis; these words are considered main influencers in the field knowledge as they could give important information in each of their occurrences (Main Words) for instance in the financial domain main words could be *Investment, Capital, Profit, Loss, etc*.

The system will first calculate each of the main words absolute frequency in a given domain related corpus (*Afreq*). The next step is calculating the frequencies of the words that come in the same segment/sentence see Fig.2 which show the segment components that are *words: w, candidate words: CW, and main word: MW* could come in random order and Fig.3 for an example that means in English (*and the value of the investment of XYZ company is about number Egyptian pound*) clarifies possible candidate words for a min word *Investment* are *value, number and Egyptian pound*.

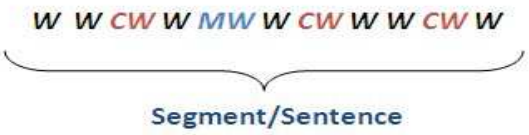

Fig. 2 Text Segment Components

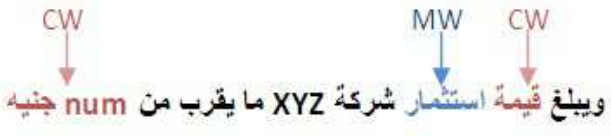

Fig. **3** Text Segment

The resulting words are candidates for relations with the main word especially those with the highest co-occurrence frequency (*Cfreq*) which calculates the times a candidate word combines the main word in the same segment/sentence in any order; it's not necessarily to come in a sequence. We can't depend only on the co-occurrence frequencies to determine that this is a strong relation; we should calculate the relation between the candidate word absolute frequency and it's co-occurrences with the main word frequency, this is can give us a better vision about how strong the relation is see Eq. (1) and (2).

$$CRfreq(MW, CW) = Cfreq(MW, CW)/Afreq(CW \quad (1)$$
$$MRfreq(MW, CW) = Cfreq(MW, CW)/Afreq(MW \quad (2)$$
$$AvgRfreq(MW, CW) = [CRfreq(MW, CW) + MRfreq(MW, CW)], \quad (3)$$
$$MAfreq(MW) = Afreq(MW)/Tokens\ numb \quad (4)$$
$$CRfreq(MW, CW1, CW2) = Cfreq(MW, CW1, CW2)/Cfreq(CW1, CW \quad (5)$$
$$MRfreq(MW, CW1, CW2) = Cfreq(MW, CW1, CW2)/Afreq(MW \quad (6)$$

Where:
Cfreq= Co-occurrences Frequency
Afreq= Absolute Frequency
CRfreq= Candidate Word Relative Frequency
MRfreq= Main Word Relative Frequency
AvgRfreq= Average Relative Frequency
MAfreq= Main word Absolute Frequency

Assume we have a main word with Absolute Frequency *Afreq(MW)* was equal to 200 times and many candidates, candidate word *Value* Co-occurrence Frequency *Cfreq(MW, CW1)* was equal to 50 times. According to the following steps after having the absolute frequency of the main word in our example *Investment* and the co-occurrences frequency of the candidate word *Value* we calculate the percentage of this frequency twice; once rated to the main word absolute frequency and another rated to the candidate word absolute frequency (Relative Frequency).

These calculations are shown in steps 3 and 4; step 3 percentage indicates the strength of the relation for the main word and step 4 percentage indicates the strength of the relation according to the whole times of the candidate word occurrences. We calculated both percentages as we can't depend on only one of them, the first percentage could be very small to consider as shown in step 3 it is 25% but when we look at this candidate word absolute occurrences we find that it mostly appears with this main word as calculated in step 4 about more than 60% of this candidate word occurrences happens to be with the main word of interest, which gives a strong indication for their relation.

In this model we considered the average of the two percentages see Eq. (3) as the determination of a relation existence which is calculated in step 5 to be 43.75%; another suggestion is putting a threshold for each percentage to consider separately.

- *Calculations Steps:*
  Afreq(MW) = 200 time.
  Cfreq(MW, CW) = 50 time.
  Main Word Relative Frequency MRfreq(MW, CW)= 50/200 = 25%
  Candidate Word Relative Frequency CRfreq(MW, CW1) = 50/80 = 62.8%
  Average Relative Frequency avgRfreq(MW, CW) = (25% + 62.8%)/2 = 43.75%

The *CRfreq* calculation could show a significant relation, for instance if the percentage was 100% this means that in all candidate word occurrences it only happened to appear with the main word in the same sentence, which represents a very strong relation to record see Fig. 4, a- sentence means in English (*the company capital is evaluated to …*) here is capital word in Arabic contains two parts each one could have a meaning independently first part means *head* and second means *money* which can be considered a main word, but when combined they form the *capital* translated word, the relation between *capital* parts words could represent the case mentioned above as the part meaning *head* could be a possible candidate to money; when we calculate the *CRfreq* percentage it often will be 100% which indicates that it only comes combined to the main word *money*, b- sentence means in English (*and the profit of XYZ Company through the period of … was*). The relation *through and period* also could represent the same case if *period* only came with *through* in this corpus, the only difference here is this relation isn't linked directly to the main word it is between two candidate words which will be elaborated in

detail in *mapping the words network* section.

a- رأس مال الشركة يقدر بنحو

b- وكانت قيمة أرباح شركة XYZ خلال الفترة

Fig. 4 a, b dependant words relation

The highest the percentage is the strongest the relation and can be linked to the word of interest. Another step is to identify the word position to the main word if it's a left hand side (LHS) or right hand side (RHS) as this can help in building more specific relation and improve the usage performance see Fig. 5. This is can be done by comparing the main word token number with the candidate word token number, if the candidate word token number is greater than the main word token number so it came after so it's a LHS and if less it's RHS in Arabic text.

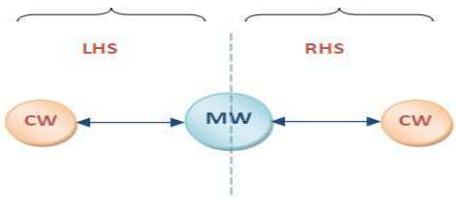

Fig. 5 Left and Right Hand Sides

B) Using system listed words:

In this case instead of feeding the system with the main words of interest we depend on the words absolute frequencies and get the highest percentages of words occurrences rated to the corpus tokens number Eq. (4) and consider them the main words of interest to search for their related words. Before considering the words with highest occurrences percentages main words a filtration step should be applied to check for meaningless words such as (prepositions, symbols, pronouns and particles), to avoid miss-categorization or non-useful relations.

The resulted percentages could be taken as proof of relating these candidate words to the main word. The system user can depend on the system fully automated relations building by setting a threshold for deciding which relations should be kept and which should be discarded or he can validate the formed relations manually with a domain expert. It's not recommended to set the threshold with a fixed value through all text corpuses as it could cause data loss due to the differences in the words occurrences frequencies according to the domain corpus used, instead different values are tested in each run till we reach the best value that keeps the important relations and avoids the weak relations.

Mapping the words network

Words can be mapped into linked nodes form to clarify the different words relations including directly linked words or indirectly linked words levels and L/RHS words (note that directly linked words not necessarily be directly sequenced to the word). Direct links are formed through the previous clarified calculations, indirect links represents another level of relations as it means that the word is dependent on another linked candidate word in other words the occurrences of this indirect word with the main word only happens with the presence of this mediate candidate word. Looking at the example shown in Fig.4.b- we explained similar case. To find these relations we should calculate the co-occurrences between the candidate words to check for any clear significant relation. Then we complete the calculations in the same logic with a difference that we calculate on the scale of the three words (if we are talking about a main word and two candidates) so the equation would take the additional candidate word in consideration Eq. (5) and (6) also we consider the co-occurrences of two candidate words is similar to absolute frequency of one candidate word in these cases. If we had more words we should consider this in the calculations and link level in the network mapping.

Establishing links is based on a pre-defined threshold for the average relative frequency percentage or candidate and main words relative frequencies. Mapping words into a network would help in making the result more informative, usable and readable for different purposes see Fig. 6.

Additional feature to decide is to allow repeating a candidate word through different levels or right/left hand sides or to permit it. For example a candidate word *Value* in Fig.4.b- could come for the same sentence with the same meaning but in different order, so considering L/RHS in building relations we will calculate each side occurrences independently or we may just let the words relations unclassified to right or left.

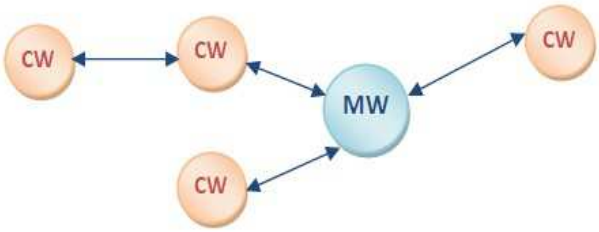

Fig. 6 Resulted Words Mapping

Updating the words network

The resulted mapped network could need to be updated through time due to corpus size and data changes, therefore the system should be able to automatically validates the strong and important links and drops useless or weak links. This could be done through the application stages (information retrieval, ontology construction, rule based matching, search, etc) using weighting approach; each time a link in the generated network is used and gets a correct result this link records a point. We should keep the data of links recorded points so in the future we can use it as a learning technique to identify the frequently used links with the highest points to keep as they represent the real life data and produce efficient results, and drop the weak links that doesn't give any results or that has very few points as it doesn't represent the real data in the domain text or it could be obsolete due to changes in words usage through time.

Another method to check for mapped words network updates is to rerun the whole system algorithm on the new formed corpus, as different corpus features could generate different words relations. A comparison between the old and new networks is held to find which relations still the same, which are new and which are dropped. The same relations kept are considered the strongest and the other relations could be filtered against threshold or via human expert validation.

Conclusion and Future work

In this paper we introduced an algorithm that can help in identifying candidate words that are considered core in a certain domain and the way to build and update relations between them. This resulted output could be used in forming information retrieval rules to represent a pattern that can get the matching important sentences also it can be used in ontology construction by extracting the main entities or updating an existing ontology with linking new unknown words.

The introduced approach hasn't been tested yet, so in the future work we intend to test this algorithm on Arabic text domain specific corpus to check its reliability and accuracy level.

References

Abdelali, A., Cowie, J., Farwell, D., Ogden, B., & Helmreich, S. (2003, June). Cross-language information retrieval using ontology. In In Proceedings of the Conference TALN 2003 (pp. 72-86).

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2010). Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet. InProceedings of The Seventh International

Conference on Language Resources and Evaluation (LREC 2010) Workshop on Language Resources and Human Language Technology for Semitic Languages (pp. 27-31).

Alanazi, S., Sharp, B., & Stanier, C. (2015). A Named Entity Recognition System Applied to Arabic Text in the Medical Domain. International Journal of Computer Science Issues (IJCSI), 12(3), 109.

Aliane, H., Alimazighi, Z., & Mazari, A. C. (2010, May). Al-Khalil: The Arabic Linguistic Ontology Project. In LREC.

Alkhalifa, M., & Rodríguez, H. (2009, May). Automatically extending NE coverage of Arabic WordNet using Wikipedia. In Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco.

Benajiba, Y., Diab, M., & Rosso, P. (2009). Arabic Named Entity Recognition: A Feature-Driven Study. IEEE Trans. Audio, Speech and Language Processing, 17(5), 926-934.

Bounhas, I., Elayeb, B., Evrard, F., & Slimani, Y. (2011). Organizing contextual knowledge for arabic text disambiguation and terminology extraction. Knowledge Organization, 38(6), 473-490.

Chang, X., Zheng, Q. (2007). Knowledge Element Extraction for Knowledge-Based Resources Organization. In InProceedings of International Conferenceon Web-Based Learning (pp. 102-113).

Jarrar, M. (2011). Building a Formal Arabic Ontology (Invited Paper). In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League.

Konswah, H. (2016). Arabic Language Textual Content Research Overview with Scope Interrelations Modeling. International Journal of Computer Science and Electronics Engineering (IJCSEE), 4(3), 167-171.

Mazari, A. C., Aliane, H., & Alimazighi, Z. (2012). Automatic Construction of Ontology from Arabic Texts. In ICWIT (pp. 193-202).

Zaidi, S., Laskri, M. T., & Bechkoum, K. (2005). A cross-language information retrieval based on an Arabic ontology in the legal domain. InProceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05) (pp. 86-91).