Automatic acquisition of word interaction patterns from corpora

Veska Noncheva
Faculty of Mathematics
and Computer Science
Plovdiv University
wesnon@pu.acad.bg

Joaquim Ferreira da Silva
Faculdade de Ciências e
Tecnologia
Universidade Nova de Lisboa
jfs@di.fct.unl.pt

Gabriel Lopes
Faculdade de Ciências e
Tecnologia
Universidade Nova de Lisboa
gpl@di.fct.unl.pt

Abstract

A major challenge in computational linguistics is to uncover word interactions in linguistic expressions. In this paper a new framework for discovering interaction between the words constituting multi-word relevant expressions is proposed. This framework is built on an algorithm for relevant expression extraction called LocalMaxs algorithm, partitioning round medoids clustering method and Bayesian networks. Bayesian networks are attractive for their ability to represent dependencies and to learn from observations. This new technology facilitates text comprehension. It may also enable control of highly ambiguous text input.

1 Introduction

Relevant expressions are sequences of text units (words, characters, part-of-speech (POS) tags, etc.) that can be extracted automatically from plain text corpora. In this work we concentrate on multi-word relevant expressions.

The relationships among words are the fundamental building blocks in the process of constituting multi-word linguistic expressions.

One of our previous aims was to cluster multi-word relevant expressions in order to separate those that denote concepts, or represent linguistic important objects (locutions, for instance), from those that clearly do not obey the above characteristics, in order to filter out the last ones (Noncheva, et al., 2002).

In this work our aim is to discover patterns of interaction between the words constituting multi word relevant expressions and to represent them by means of Bayesian networks.

By "interaction between the words constituting multi-word relevant expressions" we mean any relation holding between the lexical items, which are simultaneously presented in a single structure. For example, an interaction between the words constituting bigrams could be verb-complement relation, or the relation holding between constituents of a compound or other relations.

From such a broad definition, it appears quite clear that on principle an interaction covers a variety of linguistic phenomena ranging from collocations and compounds to selection restrictions. Information typically associated with word interaction is usually expressed in terms of semantic categories, possibly specified at different degrees of granularity (in terms of base types or of synonym clusters), and extended through use of functional restrictions. Another representational option is listing typical lexical fillers of a given word position. This kind of option could be useful when an appropriate semantic category cannot be found. Sometimes the selection involves specific lexical items rather than a general semantic class. Therefore, on the basis of the application requirements and the available linguistic information, we can decide whether to encode word restrictions through use of semantic categories, or to encode them by listing the typical lexical fillers of a given word position (this option is particularly useful when the appropriate semantic category is missing), or to combine the different encoding.

We will give a standard representation of the interaction between the words constituting multi-word relevant expressions through Bayesian networks.

The remainder of this paper is organized as follows. In Section 2 we make reference to the algorithm for extracting relevant expressions. In Section 3 we discuss the cluster approach used for extracting classes of relevant expressions. Every class contains relevant expressions that are similar in the sense that they have both the same patterns of dependence structure and similar features. In Section 4 we show how Bayesian networks are used for presenting relevant expression patterns. Bayesian models describing bigrams are presented. The experimental results are obtained from the corpus “European Legislation in force on social policy, environment, customs and rational use of energy”.

2 Extraction of Relevant Expressions

The approach LocalMaxs for finding multi-word relevant expressions from unannotated text corpora is presented in (Silva et al., 1999; Silva and Lopes, 1999). This approach is based on the idea that each *n*-gram has a kind of cohesion force sticking the word units together within the *n*-gram. In order to measure the cohesion value of each *n*-gram of any size in the corpus, a new cohesion measure using the probabilities of the *n*-grams (*n*≥1) in the corpus is proposed. As a result a data set containing thousands of relevant expressions is available. Most of them express concepts or linguistic objects that are semantically rich or technically feasible. Also, not every relevant expression denotes an interesting concept. Some examples of selected *n*-grams, are: *Human Rights*, *Human Rights in East Timor*, *common agricultural policy*, *economia energética* (in Portuguese) and *publication au Jounal officiell des Communautés* (in French).

3 Classes of Relevant Expressions

In order to find classes of Relevant Expressions, some features must be considered.

3.1 Importance measures

We would like to distinguish different written forms of words and relevant expressions. For example: *party*, *Party*, and *PARTY*.

We define the notion importance of a relevant expression extracted from a corpus in the following way:

Let *n* be the number of words w_i, *i=1,2,...,n*, in one relevant expression (*n*-gram) and L_i be the number of the different written forms of the word w_i. With w_i^* we denote the form of the word w_i in the *n*-gram under consideration. Let $\{w_i^l, l = 1,2,..., L_i\}$ be the set of all possible forms of the word w_i. Let $f(w_i^l)$ be the frequency of word form w_i^l.

We will define importance $imp_{RE}(w_1^*, w_2^*,..., w_n^*)$ of a relevant expression $w_1^*, w_2^*,..., w_n^*$ in the following way:

$$imp_{RE}(w_1^*, w_2^*,..., w_n^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{f(w_i^*)}{\sum_{l=1}^{L_i} f(w_i^l)},\text{ where}$$

$n \geq 1, L_i \geq 1, i = 1,2,...,n$.

When $n=1$ the importance of a word is

$$imp\,(w^*) = \frac{f(w^*)}{\sum_{l=1}^{L_1} f(w^l)}.$$

When $n=2$ the importance of a bigram is

$$imp_{RE}(w_1^*, w_2^*) = \frac{1}{2}\left(\frac{f(w_1^*)}{\sum_{l=1}^{L_1} f(w_1^l)} + \frac{f(w_2^*)}{\sum_{l=1}^{L_2} f(w_2^l)}\right).$$

For example, the importance of the bigrams *Mac Sharry*, *heavy-metal concentrations*, and *Fundamental Freedoms* in the corpus are:

imp_{RE}*(heavy-metal,concentration)=.99,*
imp_{RE}*(Mac,Sharry)=1,*
imp_{RE}*(Fundamental,Freedoms) =.44.*

Note that using the importance measure we will classify the bigrams *Mac Sharry* and *Fundamental Freedoms* in two clusters.

3.2 Probabilistic measures

Probabilistic information about words, phrases, and other linguistic structures is represented in the minds of language users and plays a role in language comprehension. The probability of a word’s occurrence is conditioned on many aspects of its contexts, including neighbouring words, syntactic and lexical structure, semantic expectations, and discourse factors (Jurafsky, et al., 2001).

We will focus on the role of local probabilistic relations between words in the process of acquisition of classes of relevant expressions. Consider the following measures of probabilistic relations between words:

- Prior probability

The prior probability is the probability of a word's occurrence independent of context. The prior probability is usually estimated by using the word frequency $f(w_i)$ in a sufficiently large corpus in the following way: $\tilde{p}(w_i) = \frac{f(w_i)}{N}$, where N is the total number of word tokens in the corpus. The relative frequencies $\tilde{p}(w_i)$ are estimates of the prior probabilities $p(w_i)$. The most frequent first words of the extracted bigrams $w_i w_{i+1}$ are the function words *the*, *of*, *to*, *in*, *a*, *for*, *be*, *shall*, *by*, and *on*. Their frequencies $f(w_i)$ in the corpus available are as follows: f('the')= 142222, f('of')= 100294, f('to')= 49984, f('in')= 41293, f('a')= 22821, f('for')= 21146, f('be')= 20414, f('shall')= 17234, f('by')= 15519, f('on')= 13682. The set of the most frequent second words consists of the same function words and conjunctions *and* and *or* with frequencies f('and')= 47078, f('or')= 15666.

The content words exhibit weaker effects of surrounding context, but strong effects of relative frequency (Jurafsky, et al., 2001). The frequencies of the most frequent seven content words in the corpus are as follows f('Article') = 13413, f('Member') = 11006, f('Commission') = 7457, f('Directive') = 7452, f('States') = 7081, f('Council') = 6130, f('State') = 5431.

- Joint probability

The joint probability $p(w_{i-1}w_i)$ of the bigram $w_{i-1}w_i$ may be thought of as the prior probability of these two words taken together in the same order, and could be estimated by relative frequency of the bigram: $\tilde{p}(w_{i-1}w_i) = \frac{f(w_{i-1}w_i)}{N}$, where N is the total number of bigrams.

The extracted bigrams with the highest joint probabilities are *of the*, *to the*, *in the*. They are clustered in the following two clusters {*of the*} and {*to the*, *in the*}. Other bigrams with high joint probabilities are *at least*, *in particular*, *Member States*, *Member State*, *the Commission*, *Council Directive*, and *European Communities*.

Extracted bigrams with low joint probabilities are *either voted*, *the Bank's*, *the Group's*, *not weaken*, *their non-degradability*, *Commission undertook*, *its middle*, *other media*.

- Conditional probability given previous word $p(w_i|w_{i-1})$

It is estimated from $\tilde{p}(w_i \mid w_{i-1}) = \frac{f(w_{i-1}w_i)}{f(w_{i-1})}$.

The conditional probability $p(w_i|w_{i-1})$ would be high if the second word was particularly likely to follow the first.

We have received high conditional probabilities of target words given previous word $p(w_i|w_{i-1})$ of the bigrams *Eastern Europe*, *Canary Islands*, *aquatic environment*, *lifelong learning*, and *low-skilled workers*, and low ones of *jobseekers and*, *by the*, *with the*, *failing this*, and *paternity or*.

According to the results presented in (Jurafsky, et al., 2001) the function words *of* and *to* are most likely to collocate with the previous word. For example, *kind of*, *able to*.

Our experimental results show high conditional probabilities of target function words given previous words $p(w_i|w_{i-1})$ of the bigrams *demountable as*, *insofar as*, *chaired by*, *emanating from*, and *annexed to*.

- Conditional probability given next word $p(w_i|w_{i+1})$

It is estimated from $\tilde{p}(w_i \mid w_{i+1}) = \frac{f(w_i w_{i+1})}{f(w_{i+1})}$.

We have received high conditional probability of target given next word $p(w_i|w_{i+1})$ of the bigrams *title 'Mester'*, *ISO 9001*, *global warming*, *be reconciled*, *a bunk*, and low ones of *that country's*, *this island*, *the World*, *all inventories*, and *this resolution*. We expect low conditional probabilities of target function words given next words $p(w_i|w_{i+1})$. For example "the old", "you must", "and filter". If these probabilities are high probably these bigrams are parts of n-grams, n>2.

According to the results presented in (Jurafsky, et al., 2001) the function words *I*, *a*, *the*, and *in* tend to collocate with the followings words. For example, *a lot*, *the same*, *in terms*, *I mean*.

- Conditional probability given surrounding words $p(w_i|w_{i-1},w_{i+1})$

The words w_i and w_{i-1} are content words that appear frequently and are likely to collocate with each other.

Examples: *Ebro Valley, Pergamon Press, NON-CALCAREOUS SAND, DIRECTIVITY INDEX, Onboard Observer, Merchant Shipping, Time-frame Actors, EAGGF Guidance, TOWER CRANES, White Paper, Structural Funds, Atomic Energy.*

- A class with noun phrases having property:

The first word is not predictable from the second and vice versa.

Examples: *economic viability, economic reforms, economic feasibility, economic optimum, economic restructuring, appropriate modifications, appropriate fashion, social acceptability, social forces, social spheres, new constructions, new apparatus.*

We have shown that probabilistic measures influence both extraction and clustering of relevant expressions. Now we will apply a probabilistic model called Bayesian Network for presenting and analyzing classes with relevant expressions.

4 Relevant Expression Pattern Structure

Groups of relevant expressions that present equivalent patterns have been extracted. A more ambitious goal for analysis is revealing the pattern structure.

This clearly is a hard problem. Mainly since relevant expression data alone give only a partial structure that does not always reflect key linguistic events. In addition, there is a noise in data.

In this work, we introduce a new approach for analyzing relevant expression patterns by examining statistical properties of conditional (in)dependence in the data.

The Bayesian network model is a promising tool for analyzing relevant expression patterns. First, they are particularly useful for describing processes composed of locally interacting components. Second, algorithms for inference- and learning- Bayesian networks are developed and have been used successfully in many applications. Finally, Bayesian networks provide models of causal influence (Pearl, 2000).

4.1 Bayesian Networks Language

Let χi be a set of words. By X_i we denote the word i of an n-gram, $i=1,2,...,n$. We say that X_i is a random variable with the set of its possible values χi. We can represent the dependencies between words $X_1X_2...X_n$ using a graph, in which each variable is denoted by a node. When two variables are dependent we draw an edge between them.

Consider the variables X_1, X_2 and X_3. If X_1 does not directly affect X_3, then we should say that the effect of word X_1 on word X_3 is mediated through word X_2. Once we know the word X_2, the word X_1 does not give new information about word $X_{3.}$ We formalize this in the following way: $P(X_1|X_2,X_3)=P(X_1|X_2)$ or $P(X_3|X_1,X_2)=P(X_3|X_2)$. In this case we say that X_1 and X_3 are conditionally independent, given X_2. We expect that once X_2 is fixed we will observe that X_1 and X_3 are independent. In the graph representation at Figure 1 there is not an edge between X_1 and X_3, and the relation between them is represented as a direct path through X_2.

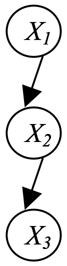
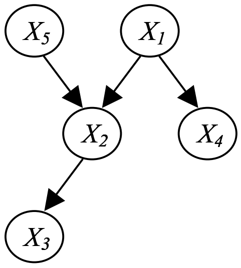


Figure 1. A simple Bayesian network structure

Consider Figure 2. As before, the three pairs of words X_1X_2, X_2X_3 and X_1X_3 are correlated. Words X_2 and X_4 are independent once we know the word X_1. Thus, word X_1 explains the relation between X_2 and X_4. In such a situation, we say that word X_1 is a common cause of words X_2 and X_4. We model graphically this relation as shown in Figure 2. If the word X_1 is not known, then X_2 and X_4 would appear dependent in data and we would have drawn an edge between them. In



$$p(x_1,x_2,x_3,x_4,x_5)=p(x_1)p(x_2|x_1,x_5)p(x_3|x_2)p(x_4|x_1)p(x_5)$$

Figure 2. An example of a Bayesian network

such a case we call X_1 a hidden common cause.

In addition to a graph that describes dependencies between variables, we associate with each variable X_i a conditional probability model that specifies the probability of X_i given its parents $\pi_i(X_i)$. We denote this probability as $p(x_i|\pi_i)$, where x_i is a value of X_i, π_i are the values of its parents.

A Bayesian network is a pair (D,P), where D is a directed acyclic graph, P={$p(x_1|\pi_1)$, $p(x_2|\pi_2)$,…, $p(x_n|\pi_n)$} is a set of n conditional probability distributions, one for each variable, and π_i are the values of parents of node X_i in D. Then the set P defines the associated joint probability distribution as $p(x_1,x_2,...,x_n)=\prod_{i=1}^{n}p(x_i\,|\,\pi_i)$. (see Figure 2).

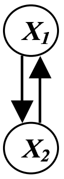
Using stochastic models is natural in word interaction because of two main reasons: the linguistic processes are stochastic and the data are noisy.

In the model described above, the sets of variables' values were sets of words. Let us note that the set of the possible values of a variable could also be a set of tags.

4.2 Bayesian networks describing bigrams

Bayesian networks describing the bigrams are discussed below. Data from the corpus have been used to learn the probabilities assigned to patterns.

Pattern 1 of dependency:



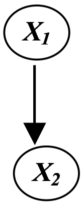
Probabilistic model:
The directed graph is a directed cycle $X_1 \to X_2, X_2 \to X_1$, representing mutual causation.

χ^1- content words: proper names, lists of nouns, lists of verbs, lists of adjectives, lists of specific words.

χ^2- content words: surnames, lists of nouns, lists of adverbs, lists of specific words.

Types of linguistic expressions:
proper names, terminologies, idiomatical verbs.
Examples: Ferro Rodrigues, Solbes Mira, Bernhard, FRIEDMANN, PINTO PIZARRO, Arctic Oceans, Bahamas Bermuda, Grand Duchy, Great Britain, United Kingdom, cathode ray, mega-electron volt, prima facie, tetra acetate, photochemical oxidants, suture stapler, bituminous shale, speech therapist, explosive atmospheres, vinyl chloride, arbitral tribunal, laid down

Pattern 2 of dependency:



Probabilistic model:

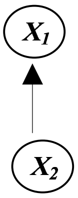
$$p(x_1,x_2) = p(x_1)p(x_2|x_1)$$

χ^1- content words: proper names, lists of nouns, lists of verbs, lists of adjectives, list of specific words.

χ^2- lists of nouns, verb forms and prepositions.

Types of linguistic expressions:
Basic grammatical forms (*it is, must be, they are*) and subcategorisation patterns (*arrange for, published in*), and terminologies.
Examples: *Coliform bacteria, Milk-based beverages, Q-L relationships, Arable crops, Labrador coast, Corinth Canal, Tarapoto Process, UKAEA Windscale, READY-TO-USE PAINTS, Devon Island, Nicosia Charter, Ross Sea, whale products, vertebrate animals, verifiable criteria.*

Pattern 3 of dependency:



Probabilistic model:

$$p(x_1,x_2) = p(x_2)p(x_1|x_2)$$

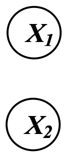
χ^1- lists of adjectives, verb forms and prepositions, and lists of specific words.

χ^2- lists of nouns, verbs and lists of specific words.

Types of linguistic expressions:
Basic grammatical patterns (*shall send*, *be elaborated*), locutions (*in particular*, *for example*),

proper names and terminologies.
Examples: ***World War, Ozone Layer, Wild Fauna, guide dogs, sensitive detector, telecommunications terminal, sugar beet, German Democratic, ISIC Nomenclature, bare ropes, traffic arteries, social affairs, commercial whaling.***

Pattern 4 of dependency:



Probabilistic model:

$$p(x_1,x_2) = p(x_1)p(x_2)$$

χ^1- lists of adjectives, verb forms and prepositions, and lists of specific words.
χ^2- lists of nouns, adjectives, verb forms and lists of specific words.
Types of linguistic expressions:
Noun and adjective phrases, and the most frequent bigrams: locutions (***as regards, at least***), basic grammatical forms of verbs (***has been, are not***), noun phrases introducing the main text topics (***European Communities***).
Examples: ***as regards, at least, has been, are not, European Communities, appropriate modifications, appropriate fashion, social acceptability, social forces, social spheres, new constructions ,new apparatus, national focal.***

The bigram class extraction result is surprisingly accurate. We are currently extending this approach in studying *n*-grams (*n*>2). When *n*>2 we expect each variable to depend on just a small subset of other variables. It will permit us to decompose the joint probability distribution function of an *n*-gram into several distributions involving a small subset of variables and then to piece them together coherently to answer question of global nature.

5 Related Works and Conclusions

This work is an alternative approach to the language-dependent extractors based on morphosyntactic filters such as Xtract (Smadja, 1990), ACABIT (Daille, 1994), etc. We have applied a probabilistic model called Bayesian Network for presenting and analyzing classes with relevant expressions. As a result certain patterns of dependency have been extracted. They reveal fragments of the underlying linguistic structures. Data from corpora are used to estimate both the dependency structures and the probabilities assigned to those structures.

This language modelling approach could be built into text entry methods in order to support the user in text production.

References

Jurafsky Daniel, Alan Bell, Michelle Gregory, and William D. Raymond. 2001. *Probabilistic relations between words: Evidence from reduction in lexical production.* In Bybee, Joan and Paul Hopper (eds.). Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins, 229-254.

Kaufman L. and P.J. Rousseeuw. 1990. Finding Groups in Data: An Introduction in Cluster Analysis, Wiley, New York.

Noncheva V., J.F Silva.. and G.P Lopes. Clustering Automatically Extracted Relevant Expressions, Technical report: CENTRIA, DI- Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal.

Pearl J. 2000. *Causality*. Cambridge University Press, Cambridge, UK.

Silva, J.F., Gael Dias, Sylvie Guilloré, and José Gabriel P. Lopes. 1999. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: P. Barahona *(ed.) Progress in Artificial Intelligence: 9th Portuguese Conference on AI, EPIA'93, Évora Portugal, September 1999, Proceedings. Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol. 1695, p. 113-132 (1999).*

Silva. J.F. and J.G.P.Lopes. 1999. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. In *Proceedings of the Sixth Meeting on Mathematics of Language* (MOL6), Orlando, Florida July 23-25, pp. 369-381.

Smadja F. and K. McKeown. 1990. Automatically Extracting and Representing Collocations for Language Generation, *Proceedings of the 28th annual meeting of the ACL*, Pittsburgh, PA.

Daile B. 1994. Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques. PhD dissertation, Universite Paris.