In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 83–86, Lisbon, Portugal, 2000.

Experiments on Unsupervised Learning for Extracting Relevant Fragments from Spoken Dialog Corpus

Konstantin Biatov
AT&T Bell-Labs Research
180 Park Avenue
07932-0971, Florham Park, NJ, USA,
Kbiatov@aol.com

Abstract

In this paper are described experiments on unsupervised learning of the domain lexicon and relevant phrase fragments from a dialog corpus. Suggested approach is based on using domain independent words for chunking and using semantical predictional power of such words for clustering and automatic extraction phrase fragments relevant to dialog topics.

1 Introduction

We are interested in rapid development of spoken dialog understanding systems. We present experiments on unsupervised learning of the domain lexicon and relevant phrase fragments from dialog corpus.

Pereira (1993) described a method for automatically clustering words according to their distribution in particular syntactic context, for example verbs and direct objects of these verbs.

By using preexisting concepts from Wordnet database Resnik (1998) described how to predict words meaning from their distributional context. Both mentioned methods are fully unsupervised and are focused only on following word distribution. They describe the dependences between verb and noun as a direct object of the verb. A new method for gathering phrases into clusters was described by Arai (1999). This method uses following and preceding words distribution and call-types, associated with each utterance, but requires at the beginning labeling and transcribing a small number of the utterances.

In contrast with the mentioned methodologies, we are interested in finding a limited set of domain independent words (less than 1000) including prepositions, adverbs and adjectives and using these words for unsupervised clustering and automatic extraction of the relevant knowledge from dialog corpus.

2 Description of the algorithm

There are four main steps in our approach.

First step is to make automatically labeling and to chunk each sentence from spoken dialog corpus into a set of short subphrases. We assume that in the spoken dialog a sentence consists of slightly related subphrases. In our experiment for labeling and chunking we use a relatively small set of domain independent words such as prepositions, determiners, articles, modals and adverbs. For example articles: **a**, **an**, **the**; prepositions: **in**, **with**, **about**, **under**, **for**, **of**, **to**; determiners: **some**, **many**. The domain independent words are grouped in subvocabularies. For instance, subvocabulary <*article*> includes words **a**, **an**, **the**. Some subvocabularies include only one word. If a given sentence includes article **A** we'll replace it by the label (<*article*>**A**), article **THE** we'll replace by the label (<*article*>**THE**) and so on. Very important feature of our algorithm is that some of the words selected for tags can predict the semantics of the followed words or subphrases. In all cases we could characterize this prediction as possibility. For example the word **from** predict semantics of the followed words or subphrases as a "start point", a "reason", a "source" or something else. For each of such tag words we create separate subvocabulary. In the process of labeling we examine given sentence from left to right and replace the tag words by the labels. For labeling we use tools based on AT&T CHRONUS system described by Levin (1995).

In the process of chunking we examine the sentence from left to right. In one chunk we put one tag word label or tag word labels following one by one and other non tag words on the right up to but excluding next tag word label. There are two examples of the chunks:

(<*what*>**WHAT**) TYPE,
(<*pronouns*>**I**) (<*article*>**A**) FARE.

We'll describe each non tag word by the vector of the features. Every component of the vector corresponds to subvocabulary of the tag words as it is described below:

component 1	→	(<*article*>**...**)
component 2	→	(<*determiner*>**...**)
component 3	→	(<*modal*>**...**)
component 4	→	(<*of*>**OF**)
component 5	→	(<*to*>**TO**)
component n	→	(<*from*>**FROM**)

Every component mean how many times tag word label was in the left context of described non tag word. Every component is an integer. Thus we have the list of non tag words and vectors of integers corresponding to this words.

Second step is to cluster the words from all chunks by using the vectors of the features. In this step we extract from chunks the words which have enough semantically charged tags in the left context and group such words in the clusters.

For clustering we take from the list the first non tag word and check if the number of different tags (number of non zero components of the vector) is more then threshold. The threshold value must be greater then the number of tag words having low semantical predictional power (articles, modals, auxiliaries, determiners). In our experiments we used threshold values from 6 up to 9. If the number of different tags for tested vector is more than threshold we'll consider this vector as a centre of the cluster and then looking for other vectors neighbouring to tested vector. When the neighbouring vectors are selected we'll remove them from the list of vectors. This procedure we'll repeat for all vectors non selected as a member of the class. For this experiments we have used distance measure based on Hamming metrics.

In the third step we go back to the chunks and extract chunks which include words from one cluster. In this way we generate the clusters of the chunks.

In the forth step we reduce the number of the chunk's clusters. We make union of all chunk's clusters except one tested cluster and then intersect this one with chunk's union. If all chunks from tested cluster are inside of the union we delete this tested chink cluster.

Let us consider baseline algorithm which use "stop words" known in information retrieval systems. The idea of this algorithm is to delete the stop words from given sentence and return all of the remaining words as lexicon items.

There are some principal differences between baseline algorithm and suggested algorithm. In suggested algorithm we are looking for the words which have enough semantically charged tags in the left context and then extract chunks which include selected words. In the baseline algorithm we are looking for only words remaining after deleting "stop words".

3 The results of experiments

Below we show examples of labeling and chunking the phrases. As an example we use two phrase from ATIS dialog corpus which includes nearly 20K sentences about flights, reservations, tickets, prices, car rent, flight classes and others.

WHAT TYPE OF AIRCRAFT IS USED FOR THIS FLIGHT

IS A MEAL SERVED FOR THIS FLIGHT

After labeling we'll have followed labeled phrase:

(<*what*>**WHAT**) TYPE (<*of*>**OF**)
AIRCRAFT (<*auxiliary*>**IS**)
USED (<*for*>**FOR**)
(<*determiner*>**THIS**) FLIGHT

(<*auxiliary*>**IS**) (<*article*>**A**)
MEAL SERVED (<*for*>**FOR**)
(<*determiner*>**THIS**) FLIGHT

In one chunk we put the tag word label or sequence of tag word labels from the left context and other non tag words on the right up but exclude the next tag word label. Below is the list of the chunks for those two sentences.

(<*what*>**WHAT**) TYPE
(<*of*>**OF**) AIRCRAFT

(<*auxiliary*>**IS**) USED
(<*for*>**FOR**) (<*determiner*>**THIS**) FLIGHT
(<*auxiliary*>**IS**)
(<*article*>**A**) MEAL SERVED

We have divided the corpus into two parts. For each part we did the labeling, chunking and clustering by using Hamming metrics for distance measure. Below we present the words extracted from both parts of the corpus.

The words extracted from the first 15K sentences.

> AIRLINE, AIRLINES, AIRPORT, AVAILABLE, BREAKFAST, BUSINESS, CITY, CLASS, COACH, COST, DAY, DINNER, EACH, EARLIEST, EARLY, ECONOMY, FARE, FARES, FLIGHT, FLIGHTS, FLY, FLYING, GET, GO, GOING, GROUND, INFORMATION, LATEST, LESS, LUNCH, MAKE, MEAL, MEALS, MOST, NONSTOP, CLASS, NUMBER, OTHER, PLANE, PRICE, RENTAL, RESTRICTIONS, RETURN, ROUND, SERVE, SERVED, SERVICE, SHOW, STOP, STOPS, TAKE, TIME, TRANSPORTATION, TRIP.

There are 54 words. Near 80% of the words could be considered as having strict relations to the dialog topics. There are such words as AIRLINE, CLASS, COACH, COST, MEALS. To understand does this approach is robust we applied the same methodology for the last 5K sentences of the corpus. The words extracted in this experiments are:

> AIRPORT, FLY, FLIGHT, STOPOVER, INFORMATION, FLIGHTS, AIRCRAFT, CITIES, COST, LESS, FARE, TRIP, ROUND, TRANSPORTATION, GROUND, LIST, FARES, CAR, TIMES, NUMBER, NONSTOP, AIRLINE, MEALS. AVAILABLE, AIRLINES.

With exception of STOPOVER, CITIES, TIMES all other words are among those extracted from first 15K sentences.

Below we present as an example the contents of the chunk's cluster for word COST extracted from the first 15K sentences:

> LOWEST COST FARE, LIKE COST, FLIGHT COST, KNOW COST, FLIGHTS COST, LOVEST COST AIRFARE, LIMOUSINE COST, RENTAL CAR COST, LIMOUSINE SERVICE COST, WITHIN CITY, NEED COST, LOWEST COST FARE ORIGIN_CITY, CHEAPEST COST FARE DEST_CITY, AVERAGE COST, AIRPORT TRANSPORTATION COST, CAR RENTAL COST, TAXI COST, MID SIZE CAR COST, ROUND TRIP COST CITY, ROUND TRIP COST, FARES COST, SEE TOTAL COST, SHOW COSTS, GIVE APPROXIMATE COST, AIR TAXI COST, COST GET, TOTAL COST, SAME COST, COST FLY, COST LESS, COST TRAVEL ORIG_AIRP, COACH COST, COST ASSOCIATED, ECONOMY ROUND TRIP TICKET COST ORIGIN_CITY, DESCENDING COST, FARE CODE F COST, COST TAKING COACH ORIGIN_CITY

And from 5K last sentences:

> AIR TAXI COST, COST, COST LESS, SEE COST, ROUND TRIP COST, COST ASSOCIATED, ECONOMY ROUND TRIP TICKET COST, DESCENDING COST, FLIGHT COST FARE CODE F COST, COACH FARE COST, COST TAKING COACH ORIGIN_CITY, COST NUMBER, CAR RENTAL COST, COST INFORMATION, COST NUMBER, FLIGHT COST LESS, COST COACH FARE ROUND TRIP TICKET, WHOSE COST, LEAST EXPENSIVE COST, ECONOMY CLASS COST.

4 Conclusion

The experiments show that the suggested method gives robust results for relevant knowledge extraction from dialog corpus.

References

K. Arai, J. Wright, G. Riccardi, and A. Gorin. 1999. Grammar fragments aquisition using syntactic and semantic clustering. *Speech Communication*, 27:43–62.

E. Levin and R. Pieraccini. 1995. Chronus, the next generation. In *Proceedings of 1995 ARPA Spoken Language Systems Technical Workshop*. Austin, Texas.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, pages 183–190. Association for Computitional Linguistics.

P. Resnik. 1998. Wordnet and class-based probabilities. In C. Fellbaum, editor, *WORDNET An electonic lexical database*, pages 239–263. The MIT Press, Cambridge, Massachusetts, London, England.