IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 6, Ver. IV (Nov.- Dec. 2017), PP 01-06
www.iosrjournals.org

A Factual Research of Word Class-Based Features For Natural Language Processing

M.VeeraKumari [1], Prof.B.Prajna [2]
[1](Research Scholar,Computer Science and Systems Engineering , Andhra University,Visakhapatnam,India)
[2](Professor,Computer Science and Systems Engineering , Andhra University,Visakhapatnam,India)
Corresponding Author: M.VeeraKumari

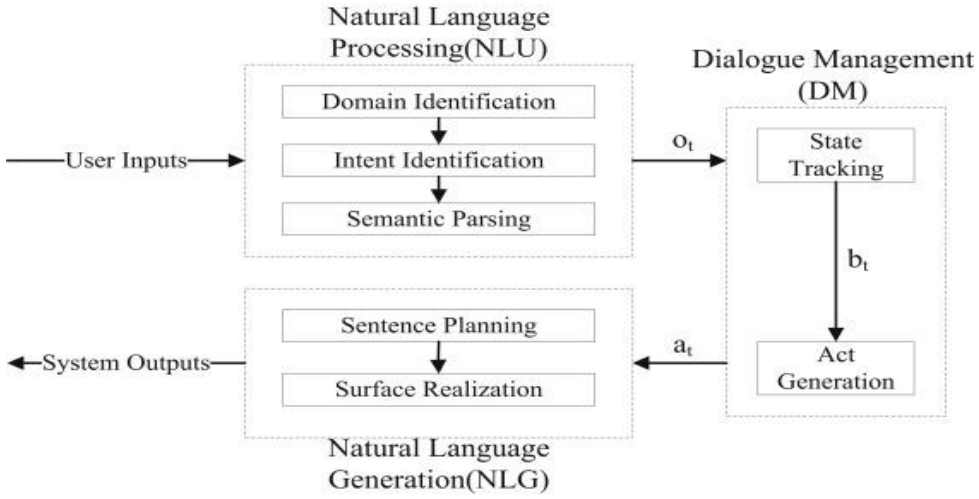
Abstract: *Using word class-based features improves the performance of natural language processing tasks based on factual observation that decline the sum of parameter magnitudes. In this paper, we explore the sequel of the word class-based features focusing on NLU tasks and indicate that the performance improvements could be attributed to the standardize effect of the class-based features. We show that class-based features extracted from different data sources using alternate word clustering methods can individually impart to the performance gain. We analyzed the actual basis of features improve the model accuracy and showed the connection with shrinkage in the model size. Since the proposed features are generated in validation of task independence on different classification and sequence tagging tasks .*

Keywords: *Natural Language Understanding, Standardize, Sequence Tagging Tasks, Word Class-Based Features.*

Date of Submission: 27-11-2017 Date of acceptance: 28-12-2017

I. Introduction

Natural Language Processing refers to the study and development of computer systems that can interpret speech and text as humans naturally speak and type it. we all use colloquialisms, abbreviations, and don't bother to correct misspellings (especially on the internet). These inconsistencies make computer analysis of natural language difficult at best, but in the last decade NLP as a field has progressed immeasurably. The information could be at the sentence level (e.g. domain detection, user intent detection) or at the word/phrase level (e.g. semantic concepts, entities). The latter combined with the former provides a granular understanding of the user's goal and also allows formulating queries to fetch information from the knowledge back-end for these applications. The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands[15]. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context. Challenges in natural language processing frequently involve speech recognition, natural language understanding, natural language generation (frequently from formal, machine-readable logical forms), connecting language and machine perception, dialog systems.



1.1 Speech Recognition

Given a sound clip of a person or people speaking, determine the textual representation of the speech[4]. This is the opposite of text to speech and is one of the extremely difficult problems colloquially termed "AI-complete". In natural speech there are hardly any pauses between successive words, and thus speech segmentation is a necessary subtask of speech recognition. Note also that in most spoken languages, the sounds representing successive letters blend into each other in a process termed co-articulation, so the conversion of the analog signal to discrete characters can be a very difficult process.

II. Overview of Related Work

The authors P. Xu and R. Sarikaya proposed, "Joint intent detection and slot filling with convolutional neural networks [1]," the key learnings and its features are automatically extracted through CNN layers and shared by the intent model. To evaluate the slot filling model, we use the ATIS corpus. We mainly compare with the three recently introduced NN based slot filling models . While it is not straightforward for these models to simultaneously handle intent classification, they all produced the new state-of-the-art slot filling results in the literature. A. Deoras and R. Sarikaya studies [2] deep belief network based semantic taggers for language model and the state-of-the-art approaches for slot filling among others use discriminative statistical models, such as conditional random fields(CRFs) for modeling. Slot filling is framed as a sequence classification problem to obtain the most probable slot sequence given some word sequence. We used word confusions to improve various spoken language understanding tasks in a CRF framework.S. F. Chen proposed[4], shrinkage-based gains will decrease as training sets increase in size, we still find significant gains even on tasks where over a billion words of training data are available. we evaluate several methods for data/model combination with Model M and rMDI models on limited-scale domains, to uncover which techniques should work best on large domains. rMDI models can give gains against other techniques for domain adaptation on moderately-sized corpora, it does not outperform simple linear interpolation on large data sets. In summary, despite the advances in language modeling over the past decades, word n-gram models remain the technology of choice in systems both large and small.The authors R. Sarikaya, S. F. Chen, B. Ramabhadran, proposed in the paper "Shrinkage based features for natural language call routing[3]" Joint training of intent and slot models has been investigated in the literature. A number of standard classifiers can be used for intent detection, such as logistic regression and support vector machines. For slot filling, conditional random field (CRF) is a proven technique and has been used extensively. The experimental results on two call-routing tasks show consistent gains over lexical features on small to medium training set sizes. As training data increases, the gains diminish.The authors R. Sarikaya, A. Celikyilmaz, proposed in the paper, "Shrinkage based features for slot tagging with conditional random fields[5]" a set of class-based features that are generated in an unsupervised fashion to improve slot tagging with Conditional Random Fields (CRFs). these features with CRFs and show that they consistently improve the slot tagging performance against baselines on several natural language understanding tasks. We applied a simple empirical rule for shrinking exponential models to the conditional random fields.

III. Natural Language Understanding Models

A common approach to building NLU models is the cascaded configuration, where the user utterance is run through domain detection followed by intent detection and slot filling to extract its semantic components separately. The domain detection determines the high level user intent (e.g. movies, music, games etc.) whereas the intent detection determines the precise user intent (e.g. find-movie, purchase-movie, play-movie etc.) within the domain. Slot filling extracts entities and other information bearing words/phrases needed for the application back-end. The intent detection is considered as a classification task to capture the user's intention and slot filling is considered as a sequence learning task specific to a given domain [6].

1.2.1. User Intent Detection

We chose a supervised learning method from the exponential family for the utterance intent classification task. We use *k* binary *l*2 regularized Logistic Regression (LR) (also known as Maximum Entropy) models to map each utterance to one of the pre-defined *k* intent classes. Each LR model uses a linear predictor function $f(k, i) = wk \cdot xi$ to predict the probability that observation *xi* has outcome *yk*, where *wk* is the vector of regression coefficients associated with the *k*th outcome (of total of *K* outcomes):

$$ln \frac{Pr(Yi = k)}{Pr(Y\ i = K)} = wk.xi \quad (1)$$

Exponentiating both sides helps to solve for the probabilities:

$$Pr(Yi = k) = \frac{ewk \cdot x}{\sum_{k=1}^{K-1} ewk \cdot xi} \quad (2)$$

The outcome of each model is combined into a softmax function thus serves as the equivalent of the logistic function in binary logistic regression.

2.1 **Intent Detection**

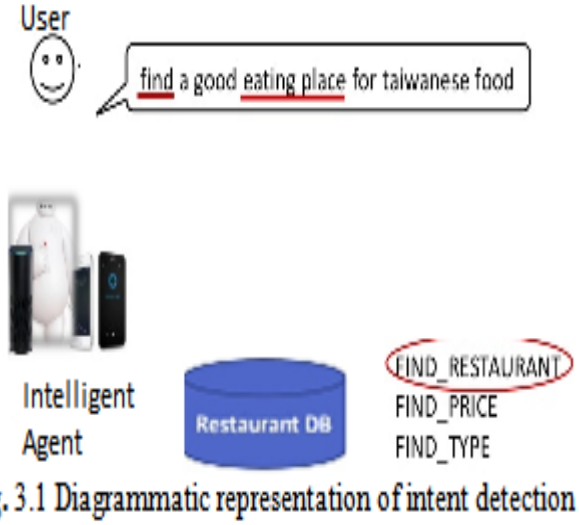


Fig. 3.1 Diagrammatic representation of intent detection

2.2 Semantic Tagging via Slot Filling

For slot filling we use conditional random fields (CRFs) [2] from the exponential family of models. CRFs are discriminative undirected probabilistic graphical models trained to maximize the conditional probability of labels on output nodes given the observations on the input nodes. If the output nodes of the graphical model are linked by edges in a linear chain, CRFs make a first-order Markov independence assumption, and thus can be understood as conditionally-trained finite state machines (FSMs). Unlike ordinary classifiers (e.g. maximum entropy models), which predict a label for a single sample without considering the neighboring labels, a CRF can take (label) context into account and model sequences of labels. For example, a linear chain CRF predicts sequences of slots for sequences of input samples (i.e. words). Assuming that *n* is the length of the observation sequence, a linear chain CRF can be written as:

$$pw(\mathbf{y}|\mathrm{x}) = \frac{1}{Zw(x)} \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} wifi\ (yj-1, yj, \mathrm{x}, j)\right) \quad (3)$$

where *j* denotes the position in the input observation sequence x = *{x*1, . . . , *xn}* andy = *{y*1, . . . , *yn}* is the output sequence. *fi*(.) are often binary valued (but can be real-valued as well) feature functions, which depend both on the input observation sequence and output label sequence. Model parameters (*wi*) are learned weights associated with feature *fi*(・) and they are independent of the position *j*. *Zw*(x) is the normalization term to make sure the expression is a probability:

$$Zw(\mathbf{x}) = \sum_{y \in Y} \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} wifi\ (yj-1, yj, \mathrm{x}, j)\right) \quad (4)$$

where summation over *Y* , the set of all possible label sequences, makes the probabilities sum to one. Within the exp(・) function, we sum over *j* = 1, . . . , *n* word positions in the sequence. Given such a model the most likely label sequence for an input sequence *x* is,

$$y^* = \arg_y \max\ Pw(\mathbf{y}|\mathbf{x}) \quad (5)$$

This expression can be efficiently computed using the Viterbi algorithm. Belonging to the same exponential family of models, CRFs share many of the properties of standard maximum entropy models, including their convex likelihood function, which guarantees that the learning procedure converges to the global maximum. Traditional maximum entropy learning algorithms, such as GIS and IIS [9], can be used to train CRFs. However it is widely observed that a stochastic gradient descent (SGD) converges much faster than GIS or IIS, so we use SGD for learning the model parameters.

2.3 **Slot Filling**

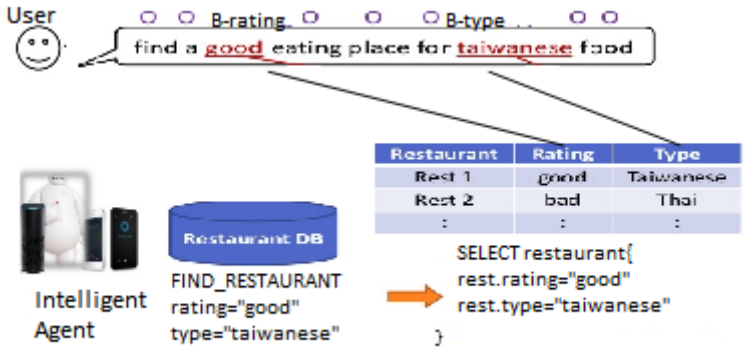


Fig. 3.2 Diagrammatic representation of slot filling

2.3 Word class-based features for shrinking parameters of the nlu models

Word class features used for shrinking the size of the NLU models, specifically multi-label LR for intent detection and linear-chain CRFs for slot filling. One common approach for inducing word classes is to use a clustering technique, preferably hierarchical. Words can be represented as discrete entities or as vectors of scalar values in high dimensional spaces through embedding. Word clustering can done using either of these representations through k-means or Brown clustering [6].

2.4 The Brown Clustering Algorithm

The Brown algorithm is a hierarchical clustering[12] algorithm which clusters words to maximize the mutual information of bigrams. So it is a class-based bigram language model. It runs in time $O(V \cdot K2)$, where V is the size of the vocabulary and K is the number of clusters. Brown clusters have been used successfully in a variety of NLP applications: NER, dependency parsing, and semantic dependency parsing.

Fig 4.1 word clustering

V is the set of all words seen in the corpus w1, w2, . . . wn
Say C : V ! {1, 2, . . . k} is a partition of the vocabulary into k classes
The model:
p(w1, w2, . . . wn) =Yn i=1 e(wi|C(wi))q(C(wi)|C(wi_1))
(note: C(w0) is a special start state)

2.5 An Example

p(w1, w2, . . . wn) =Type equation here.Yn
i=1
e(wi|C(wi))q(C(wi)|C(wi_1))
C(the) = 1, C(dog) = C(cat) = 2, C(saw) = 3
e(the|1) = 1, e(cat|2) = e(dog|2) = 0.5, e(saw|3) = 1

q(1|0) = 0.2, q(2|1) = 0.4, q(3|2) = 0.3, q(1|3) = 0.6
p(the dog saw the cat) =

2.6 A Brown clustering model consists of:
A vocabulary V
A function C : V ! {1, 2, . . . k} defining a partition of the vocabulary into k classes
A parameter e(v|c) for every v 2 V, c 2 {1 . . . k}
A parameter q(c0|c) for every c0, c 2 {1 . . . k}

2.7 Word Embeddings Method

A word embedding is a con- tinuous representation of a word. It is a mathematical object associated with each word of the vocabulary. In the literature, there are many techniques used for word embedding. Word embeddings are easy to work with because they enable efficient computation of word similarities through low-dimensional matrix operations..Among the state-of-the-art word embedding methods is the skip-gram with negative sampling model (SKIPGRAM), and implemented in the word2vec software[10]. Not only does it produce useful word representations, but it is also very efficient to train, works in an online fashion, and scales well to huge corpora (billions of words) as well as very large word and context vocabularies[13].

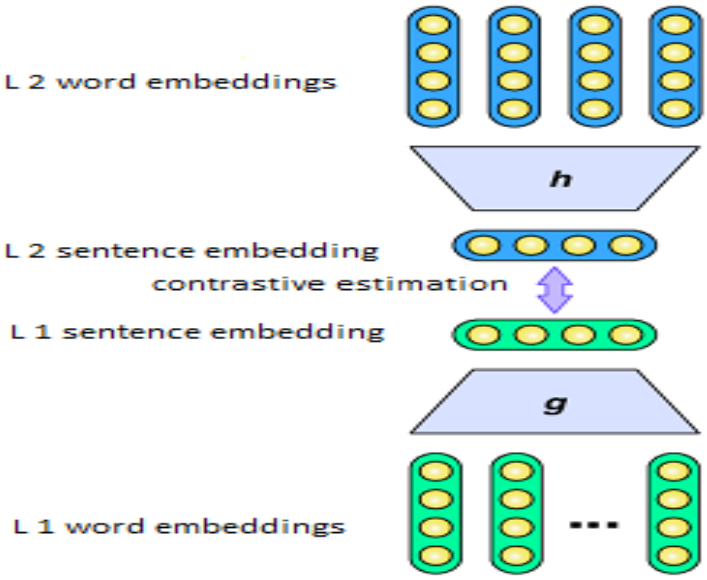


Fig 4.2 Diagrammatic representation of word embeddings

IV. Conclusion

The study in this paper found that investigated the effect of word class-based features for exponential family of models on natural language understanding (NLU) tasks. We analyzed the factual basis of why these features improve the model accuracy and showed the connection with the shrinkage in the model size. It is well known that the performance of the intent detection and slot filling tasks, which are the two core components of many NLU tasks, can be improved using additional information coming either from syntactic or semantic content of the sentence or various external resources. We are planning to investigation of the language independence to see whether these features can achieve similar gains for classification and tagging tasks in other languages.

References

[1]. P. Xu and R. Sarikaya, "Joint intent detection and slot filling with convo- lutional neural networks," Proc. IEEE ASRU, Olomouc: Czech Republic, 2013, pp. 78–83.
[2]. Asli Celikyilmaz, IEEE, Ruhi Sarikaya, IEEE, Minwoo Jeong, IEEE, and Anoop Deoras, IEEE,"An Empirical Investigation of Word Class-Based Features for Natural Language Understanding".
[3]. A. Deoras and R. Sarikaya, "Deep belief network based semantic taggers for spoken language understanding," Proc. Interspeech, Lyon: France, 2013.
[4]. R. Sarikaya, S. F. Chen, B. Ramabhadran, "Shrinkage based features for natural language call routing," Proc. Interspeech, Florence: Italy, Aug. 2011.
[5]. S. F. Chen, "Shrinking exponential language models," Proc. HLT- NAACL, Boulder: CO, USA, 2009.
[6]. R. Sarikaya, A. Celikyilmaz, A. Deoras, and M. Jeong, "Shrinkage based features for slot tagging with conditional random fields,"

Proc. Interspeech, Singapore: Sep. 2014.
[7]. J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," Proc. ACL, 2010.
[8]. G. Tur and R. De Mori, Spoken Language Understanding-Systems for Extracting Semantic Information from Speech, New York: NY, USA, Wiley, 2011.
[9]. ocher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distri- butions," Proc. EMNLP'11, 2011.
[10]. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Cavusoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., Jan. 2011.
[11]. O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N.Smith, "Improved part-of-speech tagging of online conversational text with word clusters," Proc. NAACL-HLT, Jun. 2013.
[12]. Prajna Bodapati, Evaluating the performance of a Semantic based Text Clustering Method, 2013.
[13]. Prajna Bodapati, "Potential based metrics for implementing hierarchical clustering", International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 4 Issue 8 Aug 2015, Page No. 14027-14032
[14]. G. Tur, D. Hakkani-Tur, L. Heck, and S. Parthasarathy, "Sentence Simplification for Spoken Language Understanding," in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2011, pp. 5628 –5631.
[15]. P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai and R.L. Mercer, "Class-based N-gram Models of Natural Language", Computational Linguistics, 18(4), pp: 467–479, 1992.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

* M.VeeraKumari."A Factual Research of Word Class-Based Features For Natural Language Processing." IOSR Journal of Computer Engineering (IOSR-JCE) 19.6 (2017): 01-06.