Enhancements to the Training Process of Classifier-based Speech Translator via Topic Modeling

Emil Ettelaie, Panayiotis G. Georgiou, Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory
Ming Hsieh Department of Electrical Engineering
Viterbi School of Engineering, University of Southern California
3710 S. McClintock Ave., RTH 320, Los Angeles, CA 90089, USA
ettelaie@usc.edu, georgiou@sipi.usc.edu, shri@sipi.usc.edu

Abstract

Classification of sentences based on their meaning (or concept) has been used as component in speech translation and spoken language understanding systems. Preparing training data for this type of classifiers is often a tedious task. In our previous work, we presented a method of clustering sentences as a step toward automated annotation of concepts. To measure the distance between two sentences, that method relied on the local lexical dependencies in their translations. In this work, we apply Topic Modeling to enhance the previously proposed distance metric so that it includes information from semantic associations among the words. Our experiments on the DARPA USC Transonics and BBN Transtac data sets show the advantage of incorporating this information as performance improvements in a set of clustering tasks.

Index Terms: sentence clustering, topic modeling, speech to speech translation, spoken language understanding

1. Introduction

Concept-based classification has been used effectively in Speech to Speech (S2S) translation systems as a complementary translation technique to the more conventional phrase-based Statistical Machine Translation (SMT) methods [1, 2, 3]. Concept-based classifiers are also used in a wide range of applications involving Spoken Language Understanding (SLU) [4].

The major idea behind these classifiers is that two sentences convey the same "concept" if they can be expressed with a similar translation. A significant obstacle in expanding the use of the concept-based classifiers is the cumbersome task of annotating the training data. Automatically recognizing the sentences that express similar concepts and grouping them together is an important step toward reducing the burden of annotation [1, 3]. In fact, automatic clustering of sentences in the available corpora is helpful in the annotation process. It can also lead to an unsupervised training method for classifiers.

In our previous work, we introduced a method to group utterances based on their concepts [1]. In that method, each original sentence was associated with a list of its n-best translations, generated by some available SMT engine. In spite of including some erroneous information, these n-best lists often have a more expanded lexical content than the original sentences. A language model (LM) can encapsulate the local word dependencies in each one of these lists and represent the original sentence which the list was derived from. Such LMs were used in our method, to build a table of distances between the utterances. That table was used in the clustering algorithm.

Here, by means of Topic Modeling [5] we improve the distance metric by capturing (latent) semantic associations of the words across the n-best lists. After generating n-best lists for training utterances, the topic distributions or *gist* [6] of the lists are used for distance measurements. Our experiments demonstrated improvement in clustering quality and the accuracy of the classifiers that are trained on clustered data when the topical information is included in the distance metric.

The next section gives an overview of the clustering method with n-best lists. In Section 3, the distance metrics used in our previous and current methods are explained. The details of the experiments and results are presented in Section 4 followed by conclusions in Section 5.

2. Clustering with SMT n-best Lists

For S2S translation and SLU applications, we desire to have a clustering method that groups paraphrases, i.e., sentences with common concepts, together. The conventional clustering techniques, e.g., k-means, rely merely on lexical overlap among documents, which in this case consist of only one sentence each. For instance, take the utterance *"Have a seat"*. Although it conveys the same concept as the sentences *"You may sit down"*, they do not share any lexical elements. The metric used in a conventional clustering method—usually a type of distance between word-frequency vectors—would fail to detect their similarity.

On the other hand, translating the above two sentences through a phrase-based SMT engine [7] most probably would result in n-best lists with some matching translations. Even if the generated lists do not contain any identical translations, some common words or phrases are likely to exist in both of them. The transformation of the original sentences to their n-best lists, can be viewed as a single sentence to document mapping. These n-best lists can be considered as the input documents to a clustering algorithm. The language in which these documents should be produced ("intermediate" language) can be chosen based on the availability of a high quality SMT system.

It is obvious that the quality of hypotheses in an n-best list degrades with the rank. Therefore there is an inherent trade-off between the list quality and the chance of finding matching words. In practice, the length of the lists greatly affects the performance of the whole clustering process and must be chosen carefully [1].

In summary, the proposed method is carried out in three major steps:

1. Generating n-best lists in an intermediate language via an SMT system.

28 – 31 August 2011, Florence, Italy

2. Measurement of the distance between lists with an appropriate metric.
3. Clustering based on the above distances.

The distance evaluation methods presented here and in [1] do not involve the representation of documents in a coordinate system. The adopted metrics also do not satisfy the triangular inequality. Therefore, the choice of clustering algorithms was limited to the techniques that did not rely on the above properties. Similar to [3], we used the *Exchange Method* [8] and the *Affinity Propagation* [9] for clustering. These algorithms only use the distance between elements as input. The Affinity Propagation algorithm can also operate with asymmetrical distances and therefore does not involve the symmetrization. The hierarchical class of clustering methods, e.g., agglomerative algorithms, are not suitable for these type of tasks mainly because of the chaining effect.

3. Distance Metrics

It is common in text clustering methods to represent the elements, i.e., documents or sentences, as a point in a vector space and use a metric such as Euclidean distance or cosine between the angles of two vectors as the measure of dissimilarity or similarity. However, these type of metrics are not adequately discriminative to produce the desired clusters in a concept-based fashion [1].

3.1. Language Model Distance

In our previous work, to represent each utterance in the corpus, we used an LM, built from the n-best list that SMT produced in the intermediate language. Language Models capture local word ordering. This can be viewed as capturing a shallow level of semantic and syntactic dependencies among the words that occur in the vicinity of one another throughout the corpus.

The distance between two n-best lists, and hence the dissimilarity between two original utterances, can be measured by calculating the distance between their corresponding LMs in an information theoretic fashion [10]. For instance, by using symmetric Kullback-Leibler (KL) divergence as metric and Exchange Method for clustering, we demonstrated clustering and classification performance improvements, compared to spherical k-means [1].

3.2. Topic Modeling

Although LMs are shown to be quite useful in representing the concepts of sentences, they do not explicitly model the (latent) semantic relations among words throughout documents (n-best lists, here). Topic models, on the other hand, are powerful tools to capture and manifest this sort of associations [5]. In contrast to the conventional topic modeling which relies on the word co-occurrences in a document, here we intend to learn the topics and association of words to them, based on the word co-occurrences in the multiple translations of a single sentence, i.e., an n-best list.

Here, we used Latent Dirichlet Allocation (LDA) [11]—a common method of topic modeling—to extract and represent the gist [6] of the n-best lists and, hence, the gist of the sentences that they are derived from.

The basic assumption in LDA is that the words of a document are independently generated from a set of "topics". A topic is a multinomial distribution of all the words in the corpus vocabulary. Therefore, words can be randomly generated from the topics. There is also a multinomial distribution of topics associated with each document which can be regarded as the gist of that document. Each word in a document is considered to be generated as follows: First a topic is drawn according to the gist of the document. That topic then dictates the distribution from which a word is sampled. Topic distributions (gist) are also assumed to be the samples of a Dirichlet random vector.

The topics, their distributions, and Dirichlet hyperparameters are all learned from the corpus in an unsupervised manner. The only human input beside the words is the document boundaries. Different implementations of algorithms based on either variational method [11] or Markov chain Monte Carlo (MCMC) [5] are available for training and inference.

The KL divergence[1] between the topic distributions of two n-best lists (learned via LDA), quantifies the distance between the original sentences which the n-best lists were generated from. These distances can be used for clustering, similar to the ones calculated from the LMs.

Table 1: The data sets used for clustering

Data Set	Transonics	BBN
Language	English	English
Domain	Medical	Family and Background Query
Number of classes	1,269	117
Number of Sentences	9,893	2,393

3.3. Combination of Two Distances

Inspection of the distance tables calculated based on the above two metrics revealed some level of discrepancy between them. For instance, in multiple cases, two utterances with different concepts were mistakenly gauged to be very close by one metric, while the other metric indicated their mismatch by estimating a large distance between them. Such observations induce that the metrics may contain some complementary information. This inspires seeking a method to combine the information extracted from local lexical dependencies and contained in the LM, with the information from topic modeling, learned from global semantic associations of words.

If the LM and gist of a document were statistically independent, it would be justified to simply add the distances from the two metrics. This is due to the fact that in both cases the measurements are based on KL divergence. In practice, linear combination of two distances is a simple way of incorporating information from both sources. Such combination has produced better results than using any of them exclusively, as shown in Section 4.[2]

It is noteworthy that some extensions to LDA have been introduced that can capture some level of word ordering information [12]. Our preliminary experiment with one of these methods, called *Topical N-grams*, did not lead to any promising outcome.

4. Experiments and Results

4.1. Data

To evaluate the benefits of the metric derived from topic modeling, we experimented with two data corpora shown in Table 1. Both data sets consist of multiple classes of paraphrases in English. The first set is from the *Transonics* project [1] and the

[1] Experiments with Hellinger distance showed inferior results.

[2] Adding the square of two distances, did not appear to be much beneficial in our experiments.

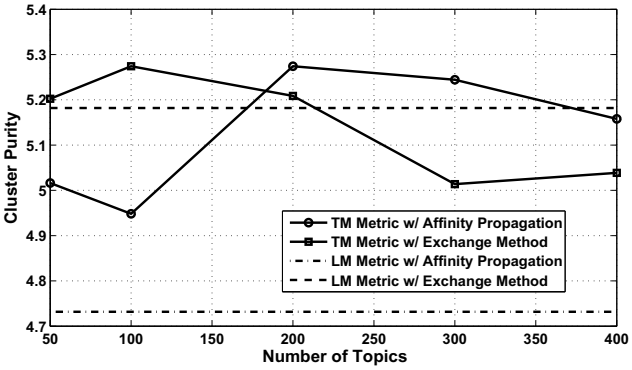

Figure 1: Purity of clustering BBN data with LM and TM metrics and different topic numbers

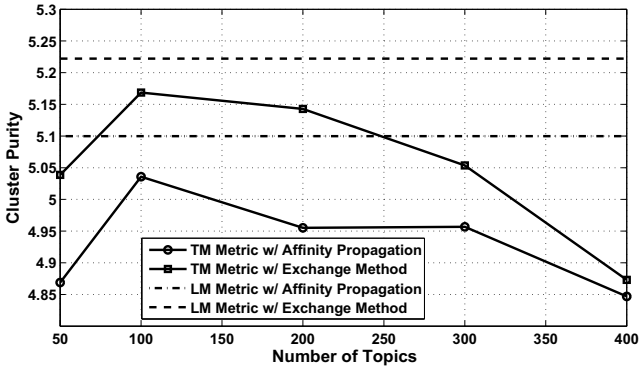

Figure 2: Purity of clustering Transonics data with LM and TM metrics and different topic numbers

second one was prepared for DARPA's *Transtac* project, and was provided to us by *BBN Technologies*.

To generate the n-best lists we trained the Moses SMT engine [7] using a parallel English/Farsi (intermediate language) corpus of size 1.18M words on the English side. The corpus was collected for the Transonics project and contains general-domain conversations and doctor-patient interactions.

4.2. Clustering with Topic Modeling Metric

The first set of experiments were intended to examine whether the metric driven from topic modeling (TM metric) can be effectively used for clustering.

For the Transonics data we only used 97 classes that have at least four utterances. We selected 500 utterances from each corpus that were randomly drawn from 97 and 117 classes of the Transonics and BBN sets respectively. The n-best lists were generated for them with the length of $n = 50$. In our previous experiments, this list size had produced relatively good results [1] and LDA models did not take much processing time to train with lists of this size.

For topic modeling, we used the *Mallet* toolkit [13] which offers the Gibbs sampling [5] method for training the LDA models. Separate models were trained for each data set through 100,000 iterations of Gibbs sampling. The Dirichlet hyperparameters were optimized every 2000 iterations. To observe the effects of different choices of number of topics, we trained models with 50, 100, 200, 300, and 400 topics. For each case a distance table was built by computing the KL divergence between the gist of every two lists.

We ran the Exchange Method and Affinity Propagation clustering algorithms on these tables. To evaluate the quality of the clustering outcome, for each case the cluster purity was calculated. If $\mathcal{R}$ is the set of reference classes, the cluster purity is defined for cluster set $\mathcal{C}$ as,

$$E = -\sum_{C\in\mathcal{C}} \frac{|C|}{|\mathcal{C}|} \sum_{R\in\mathcal{R}} P_{CR}\log(P_{CR}) \qquad (1)$$

where $|\cdot|$ is the set cardinality, and,

$$P_{CR} \triangleq \frac{|C\cap R|}{\left|C\cap\left(\bigcup_{\Theta\in\mathcal{R}}\Theta\right)\right|} \qquad (2)$$

Figure 1 shows the purity of clustering the BBN set with the TM metric, for different numbers of topics. The results from both Exchange Methods and Affinity Propagation are plotted alongside with the purity of clustering when using the LM metric. The graphs show that for some choices of the topic number, each clustering method has produced a better result with the TM metric than the result of the same method with the LM metric. Especially, this gain has been larger for the Affinity Propagation algorithm.

Similar graphs for the Transonics set are shown in Figure 2. For this data set, both algorithms produced better results with LM metric. However, for 100 topics, the purities of the resulting clusters are close to the outcome of using the LM metric. The better performance with the LM metric is due to the domain match between the Transonics data and a portion of the SMT's training set which was drawn from the medical domain. This led to better word orderings in the translations, and therefore better LMs for that Transonics data.

These experiments show that the TM metric, when used for clustering, can produce results comparable to the LM metric. For comparison of the above two algorithms using the LM metric with other clustering methods see [1].

We also used the clustered data to train a set of classifiers for each corpus. For that purpose the LM-based classification method [1] was chosen. To test the classifiers, 707 and 1,000 utterances were randomly drawn from the rest of the Transonics and BBN sets, respectively, disjoint from their training sets. Table 2 shows the concept classification accuracy results. For the cases with the TM metric, the relative change with respect to the accuracy from the LM metric is also shown. The TM metric produced better results when used with Exchange Methods. For the BBN corpus, an improvement was achieved using the TM metric, when the distance table for 200 topics (peak of the purity in Figure 1) was fed to the Affinity propagation algorithm.

4.3. Clustering with Combined Metric

In the second set of experiments a combination of both metrics were used with the Affinity Propagation algorithm which is much faster than the Exchange Method. For each data set, we combined the two distance tables from the previous experiment, one with the LM metric and the other one with the TM metric, derived with the number of topics set to 100. We scaled the

Table 2: Results of the experiments with classifiers trained on the clustered data

Corpus	Metric	Affinity Propagation		Exchange Method	
		Acc.	Rel. Acc.	Acc.	Rel. Acc.
BBN	LM	27.8%	–	31.5%	–
	TM (T=100)	26.7%	-4.0%	33.5%	6.0%
	TM (T=200)	36.5%	31.3%	–	–
Trans.	LM	49.2%	–	46.0%	–
	TM (T=100)	43.1%	-12.4%	49.2%	7.1%

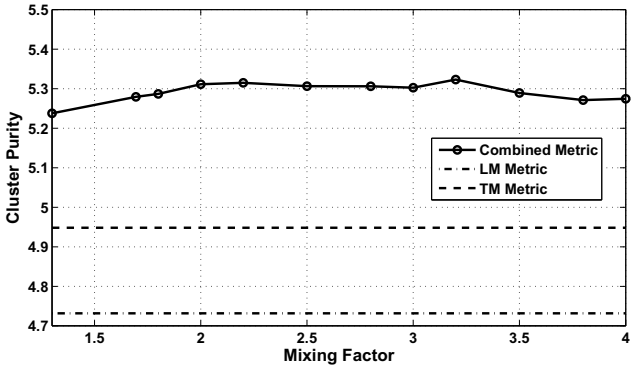

Figure 3: Purity of clustering BBN data with the combined metric

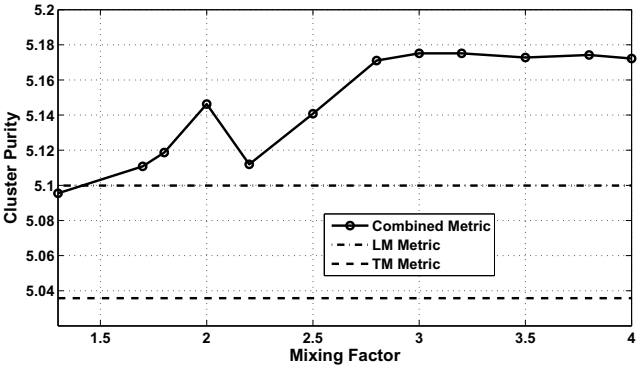

Figure 4: Purity of clustering Transonics data with the combined metric

entries of the table with the LM metric by a mixing factor and added them to their counterparts in the table with the TM metric.

Initially, to make the metrics comparable, the mixing factor was chosen as the ratio of the largest entries in the two tables. This ratio was very close for the two data sets: 1.695 for BBN and 1.699 for Transonics. The purity of the clusters produced by using the combined tables are shown in Figures 3 and 4 for the BBN and the Transonic data sets respectively. The figures show the purity for different values of the mixing factor around its initially selected value. For reference, we also included the results when using each metric exclusively. It is clear that the combined metric led to a much better clustering outcome than what the sole use of each metric had produced. For both data sets, a mixing factor of 3.2 delivered the best results. We also ran the Exchange Method for that value of mixing factor.

For a mixing factor of 3.2, we trained classifiers on the resulting clusters from both methods. For the BBN set, the classifier trained with clusters from Affinity Propagation, showed an accuracy of 39.9% which is a relative improvement of 43.5%, over the result gained from using the LM metric (Table 2). For the Exchange Method, the accuracy was 32.6%, i.e., a relative improvement of 3.5% over the baseline with the LM metric.

For the Transonic set, with the clusters from the Affinity Propagation and the combined metric, the classifier showed an accuracy of 49.4% which is only 0.3% better than what was gained from the LM metric. However, using the Exchange Method led to a relative improvement of 9.9% (again with respect to the case of using the LM metric, Table 2) as the classification accuracy reached 50.5%.

5. Conclusions

This work is the continuation of efforts towards the development of an unsupervised training method for concept-based classifiers. We have shown that by using topic modeling, the information extracted from semantic associations of words can be used to improve the quality of the sentence clustering method that we had previously introduced.

We intend to continue this work further by examining the use of other topic modeling methods and more sophisticated distance measurement techniques. We are also investigating more elaborate strategies for combining the metrics.

6. Acknowledgements

This work was supported by funds from DARPA and NSF. We would like to thank BBN Technologies for sharing their data with us.

7. References

[1] E. Ettelaie, P. G. Georgiou, and S. Narayanan, "Towards unsupervised training of the classifier-based speech translator," in *Proc. of the International Conference on Spoken Language Processing*, Brisbane, Australia, September 2008, pp. 2739–2742.

[2] F. Ehsani, J. Kinzey, D. Master, K. Sudre, D. Domingo, and H. Park, "S-MINDS 2-way speech-to-speech translation system," in *Proc. of the Medical Speech Translation Workshop, Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, New York, NY, USA, June 2006, pp. 44–45.

[3] E. Ettelaie, P. G. Georgiou, and S. Narayanan, "Unsupervised data processing for classifier-based speech translator," *ISCA Journal of Computer Speech and Language*, in press.

[4] D. Traum, A. Roque, A. Leuski, P. Georgiou, J. Gerten, B. Martinovski, S. Narayanan, S. Robinson, and A. Vaswani, "Hassan: A virtual human for tactical questioning," in *Proc. of the Eighth SIGdial workshop on Discourse and Dialogue*, Antwerp, Belgium, September 2007, pp. 75–78.

[5] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook of Latent Semantic Analysis*, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 2007.

[6] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers, "Topics in semantic representation," *Psychological Review*, vol. 114, pp. 211–244, 2007.

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, vol. Companion Proc. of the Demo and Poster Sessions, Prague, Czech Republic, June 2007, pp. 177–180.

[8] H. Sp̈ath, *The Cluster Dissection and Analysis Theory FORTRAN Programs Examples*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1985.

[9] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, February 2007.

[10] A. Sethy, S. Narayanan, and B. Ramabhadran, "Measuring convergence in language model estimation using relative entropy," in *Proc. of the Eight International Conference on Spoken Language Processing*, Jeju Island, Korea, October 2004, pp. 1057–1060.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, March 2003.

[12] X. Wang, A. Mccallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in *Proc. of the Seventh IEEE International Conference on Data Mining*, Omaha, NE, USA, October 2007, pp. 697–702.

[13] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.