Context Sense Clustering for Translation

João Casteleiro
Universidade Nova de Lisboa
Departamento de Informática
2829-516 Caparica, Portugal
casteleiroalves@gmail.com

Gabriel Lopes
Universidade Nova de Lisboa
Departamento de Informática
2829-516 Caparica, Portugal
gpl@fct.unl.pt

Joaquim Silva
Universidade Nova de Lisboa
Departamento de Informática
2829-516 Caparica, Portugal
jfs@fct.unl.pt

Extended Abstract

Word sense ambiguity is present in all words with more than one meaning in several natural languages and is a fundamental characteristic of human language. This has consequences in translation as it is necessary to find the right sense and the correct translation for each word. For this reason, the English word *fair* can mean *reasonable* or *market* such as *plant* also can mean *factory* or *herb*.

The disambiguation problem has been recognize as a major problem in natural languages processing research. Several words have several meanings or senses. The disambiguation task seeks to find out which sense of an ambiguous word is invoked in a particular use of that word. A system for automatic translation from *English* to *Portuguese* should know how to translate the word *bank* as *banco* (an institution for receiving, lending, exchanging, and safeguarding money), and as *margem* (the land alongside or sloping down to a river or lake), and also should know that the word *banana* may appear in the same context as *acerola* and that these two belongs to hyperonym *fruit*. Whenever a translation systems depends on the meaning of the text being processed, disambiguation is beneficial or even necessary. Word Sense Disambiguation is thus essentially a classification problem; given a word *X* and an inventory of possible semantic tags for that word that might be translation, we seek which tag is appropriate for each individual instance of that word in a particularly context.

In recent years research in the field has evolved in different directions. Several studies that combine clustering processes with word senses has been assessed by several. Apidianaki in (2010) presents a clustering algorithm for cross-lingual sense induction that generates bilingual semantic inventories from parallel corpora. Li and Church in (2007) state that should not be necessary to look at the entire corpus to know if two words are strongly associated or not, thus, they proposed an algorithm for efficiently computing word associations. In (Bansal et al., 2012), authors proposed an unsupervised method for clustering translations of words through point-wise mutual information, based on a monolingual and a parallel corpora. Gamallo, Agustini and Lopes presented in (2005) an unsupervised strategy to partially acquire syntactic-semantic requirements of nouns, verbs and adjectives from partially parsed monolingual text corpora. The goal is to identify clusters of similar positions by identifying the words that define their requirements extensionally. In (1991) Brown et al. described a statistical technique for assigning senses to words based on the context in which they appear. Incorporating the method in a machine translation system, they have achieved to significantly reduce translation error rate. Tufis et al. in (2004) presented a method that exploits word clustering based on automatic extraction of translation equivalents, being supported by available aligned wordnets. In (2013), Apidianaki described a system for SemEval-2013 Cross-lingual Word Sense Disambiguation task, where word senses are represented by means of translation clusters in a cross-lingual strategy.

In this article, a Sense Disambiguation approach, using Context Sense Clustering, within a mono-lingual strategy of neighbor features is proposed. We described a semi-supervised method to classify words based on clusters of contexts strongly correlated. For this purpose, we used a covariance-based correlation measure (Equation 1). Covariance (Equation 2) measure how much two random variables change together. If the values of one variable (sense *x*) mainly correspond to the values of the other variable (sense *y*), the variables tend to show similar behavior

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 135–137,
October 25, 2014, Doha, Qatar.

and the covariance is positive. In the opposite case, covariance is negative. Note that this process is computationally heavy. The system needs to compute all relations between all features of all left words. If the number of features is very large, the processing time increases proportionally.

$$Corr\ (x,y) = \frac{Cov(\ x,y\ )}{\sqrt{Cov(x,x)} + \sqrt{Cov(y,y)}} \quad (1)$$

$$Cov(x,y) = \frac{1}{m-1}\sum_{f=f1}^{fm}(dist(x,f).dist(y,f)) \quad (2)$$

Our goal is to join similar senses of the same ambiguous word in the same cluster, based on features correlation. Through the analysis of correlation data, we easily induce sense relations. In order to streamline the task of creating clusters, we opted to use *WEKA* tool (Hall et al., 2009) with *X-means* (Pelleg et al., 2000) algorithm.

Clusters
fructose, glucose
football, chess
title, appendix, annex
telephone, fax
liver, hepatic, kidney
aquatic, marine
disciplinary, infringement, criminal

Table 1. Well-formed resulting clusters

In order to determine the consistence of the obtained clusters, all of these were evaluated with *V-measure*. *V-measure* introduce two criteria presented in (Rosenberg and Hirschberg, 2007), homogeneity *(h)* and completeness *(c)*. A clustering process is considered homogeneously well-formed if all of its clusters contain only data points which are members of a single class. Comparatively, a clustering result satisfies completeness if all data points that are members of a given class are elements of the same cluster.

Analysing the results of context sense clusters obtained (Table 1) we easily understand that almost all clusters are generally well formed, getting a final *V-measure* average rating of 67%.

Finally, in order to train a classifier we choose to use a training data set with 60 well formed clusters (with V-measure value ranging between 0.9 and 1). Our testing data set is composed by 60 words related to the clusters but which are not contained there. The classifier used was a *Support Vector Machine (SVM)* (2011). The kernel type applied was the *Radial Basis Function (RBF).* This kernel non linearly maps samples into a higher dimensional space, so it can handle the case when the relation between class labels and attributes is nonlinear, that is the case. Each word of training and testing data sets were encoded according the frequency in a corpora of all characteristics contained in the clusters. Our purpose was to classify each one of the new potential ambiguous words, and fit it in the corresponding cluster (Table 2 and Table 3).

Test Words	**Label assigned by (*SVM*)**
Fruit	Cluster 29
Infectious	Cluster 7
Kiwi	Cluster 60
Back	Cluster 57
Legislative	Cluster 34
Grape	Cluster 29
Russian	Cluster 59

Table 2. Results generated by (*SVM*)

Clusters	**Content of Clusters**
Cluster 7	Viral, contagious, hepatic
Cluster 29	Banana, apple
Cluster 34	Legal, criminal, infringement
Cluster 57	Cervical, lumbar
Cluster 59	French, Italian, Belgian, German
Cluster 60	Thyroid, mammary

Table 3. Cluster correspondence

The obtained results showed that almost all words were tagged in the corresponding cluster. Evaluating system accuracy we obtained an average value of 78%, which means that from the 60 tested words, 47 words were assigned to the corresponding context cluster.

References

Marianna Apidianaki, Yifan He, et al. 2010. An algorithm for cross-lingual sense-clustering tested in a mt evaluation setting. In Proceedings of the International Workshop on Spoken Language Translation, pages 219–226.

Li, P., Church, K.W.: A sketch algorithm for estimating two-way and multi-way associations. Computational Linguistics 33 (3), 305 - 354 (2007).

Bansal, M., DeNero, J., Lin, D.: Unsupervised translation sense clustering. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 773-782. Association for Computational Linguistics (2012).

Gamallo, P., Agustini, A., Lopes, G.P.: Clustering syntactic positions with similar semantic requirements. Computational Linguistics 31(1), 107-146 (2005).

Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: Word-sense disambiguation using statistical methods. In: Proceedings of the 29th annual meeting on Association for Computational Linguistics. pp. 264-270. Association for Computational Linguistics (1991).

TufiS, D., Ion, R., Ide, N.: Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In: Proceedings of the 20th international conference on Computational Linguistics. p. 1312. Association for Computational Linguistics (2004).

Apidianaki, M.: Cross-lingual word sense disambiguation using translation sense clustering. In: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013). pp. 178-182. *SEM and NAACL (2013)

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1):10–18.

Dan Pelleg, Andrew W Moore, et al. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In ICML, pages 727–734.

Andrew Rosenberg and Julia Hirschberg. 2007. Vmeasure: A conditional entropy-based external cluster evaluation measure. In EMNLP-CoNLL, volume 7, pages 410–420.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27.