SOFTCARDINALITY: Learning to Identify Directional Cross-Lingual Entailment from Cardinalities and SMT

Sergio Jimenez, Claudia Becerra
Universidad Nacional de Colombia
Ciudad Universitaria,
edificio 453, oficina 114
Bogotá, Colombia
sgjimenezv@unal.edu.co
cjbecerrac@unal.edu.co

Alexander Gelbukh
CIC-IPN
Av. Juan Dios Bátiz, Av. Mendizábal,
Col. Nueva Industrial Vallejo
CP 07738, DF, México
gelbukh@gelbukh.com

Abstract

In this paper we describe our system submitted for evaluation in the CLTE-SemEval-2013 task, which achieved the best results in two of the four data sets, and finished third in average. This system consists of a SVM classifier with features extracted from texts (and their translations SMT) based on a cardinality function. Such function was the soft cardinality. Furthermore, this system was simplified by providing a single model for the 4 pairs of languages obtaining better (unofficial) results than separate models for each language pair. We also evaluated the use of additional circular-pivoting translations achieving results 6.14% above the best official results.

1 Introduction

The Cross-Lingual Textual Entailment (CLTE) task consists in determining the type of directional entailment (i.e. *forward*, *backward*, *bidirectional* or *no-entailment*) between a pair of texts T_1 and T_2, each one written in different languages (Negri et al., 2013). The texts and reference annotations for this task were obtained through crowdsourcing applied to simpler sub-tasks (Negri et al., 2011). CLTE has as main applications content synchronization and aggregation in different languages (Mehdad et al., 2012; Duh et al., 2013). We participated in the first evaluation of this task in 2012 (Negri et al., 2012), achieving third place on average among 29 participating systems (Jimenez et al., 2012).

Since in the CLTE task text pairs are in different languages, in our system, all comparisons made between two texts imply that one of them was written by a human and the other is a translation provided by statistical machine translation (SMT). Our approach is based on an SVM classifier (Cortes and Vapnik, 1995) whose features were cardinalities combined with similarity scores. That system was motivated by the fact that most text similarity functions are symmetric, e.g. Edit Distance (Levenshtein, 1966), longest common sub-sequence (Hirschberg, 1977), Jaro-Winkler similarity (Winkler, 1990), cosine similarity (Salton et al., 1975). Thus, the use of these functions as only resource seems counter-intuitive since CLTE task is asymmetric for the *forward* and *backward* entailment classes.

Moreover, cardinality is the central component of the resemblance coefficients such as Jaccard, Dice, overlap, etc. For instance, if T_1 and T_2 are texts represented as bag of words, it is only necessary to know the cardinalities $|T_1|$, $|T_2|$ and $|T_1 \cap T_2|$ to obtain a similarity score using a resemblance coefficient such as the Dice's coefficient (i.e. $2 \cdot |T_1 \cap T_2|/(|T_1| + |T_2|)$). Therefore, the idea is to use the individual cardinalities to enrich a set of features extracted from texts.

Cardinality gives a rough idea of the amount of information in a collection of elements (i.e. words) providing the number of different elements therein. That is, in a collection of elements whose majority are repetitions contains less information than a collection whose elements are mostly different. However, the classical sets cardinality is a rigid measure as do not take account the degree of similarity among the elements. Unlike the sets cardinality, soft cardinality (Jimenez et al., 2010) uses the similarities among the elements providing a more flexible

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 34–38, Atlanta, Georgia, June 14-15, 2013.

measurement of the amount of information in a collection. In the 2012 CLTE evaluation campaign, it was noted that the soft cardinality overcame classical cardinality in the task at hand. All the models used in our participation and proposed in this paper are based on the soft cardinality. A brief description of the soft cardinality is presented in Section 2, along with a description of the functions used to provide the similarities between words. Besides, the set of features that are derived from all pairs of texts and their cardinalities are presented in Section 3.

Section 4 provides a detailed description for each of the 4 models (one for each language pair) used to get the predictions submitted for evaluation. In Section 5 a simplified-multilingual model is tested with several word-similarity functions and circular-pivoting translations.

In sections 6 and 7 a brief discussion of the results and conclusions of our participation in this evaluation campaign are presented.

2 Soft Cardinality

The soft cardinality (Jimenez et al., 2010) of a collection of words T is calculated with the following expression:

$$|T|' = \sum_{i=1}^{n} w_i \left(\sum_{j=1}^{n} \mathbf{sim}(t_i, t_j)^p \right)^{-1} \quad (1)$$

Having $T = \{t_1, t_2, \ldots, t_n\}$; $w_i \geq 0$; $p \geq 0$; $1 > \mathbf{sim}(x, y) \geq 0$, $x \neq y$; and $\mathbf{sim}(x, x) = 1$. The parameter p controls the degree of "softness" of the cardinality (the larger the "harder"). The coefficients w_i are weights associated with each word (or term) t, which can represent the importance or informative character of each word (e.g. *idf* weights). The function **sim** is a word-similarity function. Three such functions are considered in this paper:

***Q*-grams**: each word a_i is represented as a collection of character *q*-grams (Kukich, 1992). Instead of single length *q*-grams, a combination of a range of lengths q_1 to q_2 was used. Next, a couple of words are compared with the following resemblance coefficient: $\mathbf{sim}(t_i, t_j) = \frac{|t_i \cap t_j| + bias}{\alpha \cdot \max(|t_i|, |t_j|) + (1-\alpha) \cdot \min(|t_i|, |t_j|)}$. The parameters of this word-similarity function are q_1, q_2, α and $bias$.

Group 1: basic cardinalities			
#1	$\lvert T_1\rvert'$	#4	$\lvert T_1 \cup T_2\rvert'$
#2	$\lvert T_2\rvert'$	#5	$\lvert T_1 - T_2\rvert'$
#3	$\lvert T_1 \cap T_2\rvert'$	#6	$\lvert T_2 - T_1\rvert'$
Group 2: asymmetrical ratios			
#7	$\lvert T_1 \cap T_2\rvert' / \lvert T_1\rvert'$	#8	$\lvert T_1 \cap T_2\rvert' / \lvert T_2\rvert'$
Group 3: similarity and arithmetical* scores			
#9	$\lvert T_1 \cap T_2\rvert' / \lvert T_1 \cup T_2\rvert'$	#10	$\frac{2 \cdot \lvert T_1 \cap T_2\rvert'}{\lvert T_1\rvert' + \lvert T_2\rvert'}$
#11	$\lvert T_1 \cap T_2\rvert' / \sqrt{\lvert T_1\rvert' \cdot \lvert T_2\rvert'}$	#12	$\frac{\lvert T_1 \cap T_2\rvert'}{\min[\lvert T_1\rvert', \lvert T_2\rvert']}$
#13	$\frac{\lvert T_1 \cap T_2\rvert' + \lvert T_1\rvert' + \lvert T_2\rvert'}{2 \cdot \lvert T_1\rvert' \cdot \lvert T_2\rvert'}$	#14*	$\lvert T_1\rvert' \cdot \lvert T_2\rvert'$

Table 1: Set of features derived from texts T_1 and T_2

Edit-Distance: a similarity score for a pair of words can be obtained from their Edit Distance (Levenshtein, 1966) by normalizing and converting distance to similarity with the following expression: $\mathbf{sim}(t_i, t_j) = 1 - \frac{\mathrm{EditDistance}(t_i, t_j)}{\max[\mathrm{len}(t_i), \mathrm{len}(t_j)]}$.

Jaro-Winkler: this measure is based on the Jaro (1989) similarity, which is given by this expression $\mathrm{Jaro}(t_i, t_j) = \frac{1}{3}\left(\frac{c}{\mathrm{len}(t_i)} + \frac{c}{\mathrm{len}(t_j)} + \frac{c-m}{c}\right)$, where c is the number of characters in common within a sliding window of length $\frac{\max[\mathrm{len}(t_i), \mathrm{len}(t_j)]}{2} - 1$. To avoid division by 0, when $c = 0$ then $\mathrm{Jaro}(t_i, t_j) = 0$. The number of transpositions *m* is obtained sorting the common characters according to their occurrence in each of the words and counting the number of non-matching characters. Winkler (1990) proposed an extension to this measure taking into account the common prefix length l through this expression: $\mathbf{sim}(t_i, t_j) = \mathrm{Jaro}(t_i, t_j) + \frac{l}{10}(1 - \mathrm{Jaro}(t_i, t_j))$.

3 Features from Cardinalities

For a pair of texts T_1 and T_2 represented as bags of words three basic soft cardinalities can be calculated: $|T_1|'$, $|T_2|'$ and $|T_1 \cup T_2|'$. The soft cardinality of their union is calculated using the concatenation of T_1 and T_2. More additional features can be derived from these three basic features, e.g. $|T_1 \cap T_2|' = |T_1|' + |T_2|' - |T_1 \cup T_2|'$ and $|T_1 - T_2|' = |T_1|' - |T_1 \cap T_2|'$. The complete set of features classified into three groups are shown in Table 1.

4 Submitted Runs Description

The data for the 2013 CLTE task consists of 4 data sets (*spa-eng*, *ita-eng*, *fra-eng* and *deu-eng*) each

Data set	q_1	q_2	α	$bias$
deu-eng	2	2	0.5	0.0
fra-eng	2	3	0.5	0.0
ita-eng	2	4	0.6	0.0
spa-eng	1	3	0.5	0.1

Table 2: Parameters of the q-grams word-similarity function for each language pair

with 1,000 pairs of texts for training and 500 for testing. For each pair of texts T_1 and T_2 written in two different languages, two translations are provided using the Google's translator[1]. Thus, T_1^t is a translation of T_1 into the language of T_2 and T_2^t is a translation of T_2 into the language of T_1. Using these pivoting translations, two pairs of texts can be compared: T_1 with T_2^t and T_1^t with T_2.

Then all training and testing texts and their translations were pre-processed with the following sequence of actions: *i*) text strings were tokenized, *ii*) uppercase characters are converted into lowercase equivalents, *iii*) stop words were removed, *iv*) punctuation marks were removed, and *v*) words were stemmed using the Snowball[2] multilingual stemmers provided by the NLTK Toolkit (Loper and Bird, 2002). Then every stemmed word is tagged with its *idf* weight (Jones, 2004) calculated with the complete collection of texts and translations in the same language.

Five instances of the soft cardinality are provided using 1, 2, 3, 4 and 5 as values of the parameter p. Therefore, the total number of features for each pair of texts is the multiplication of the number of features in the feature set (i.e. 14, see Table 1) by the number of soft cardinality functions (5) and by 2, corresponding to the two pairs of comparable texts. That is, $14 \times 5 \times 2 = 140$ features.

The **sim** function used was q-grams, whose parameters were adjusted for each language pair. These parameters, which are shown in Table 2, were obtained by manual exploration using the training data.

Four vector data sets for training (one for each language pair) were built by extracting the 140 features from the 1,000 training instances and using their gold-standard annotations as class attribute. Predictions for the 500 test cases were obtained through a SVM classifier trained with each data set. For the submitted *run1*, this SVM classifier used a linear kernel with its complexity parameter set to its default value $C = 1$. For the *run2*, this parameter was adjusted for each pair of languages with the following values: $C_{spa-eng} = 2.0$, $C_{ita-eng} = 1.5$, $C_{fra-eng} = 2.3$ and $C_{deu-eng} = 2.0$. The implementation of the SVM used is that which is available in WEKA v.3.6.9 (SMO) (Hall et al., 2009). Official results for *run1*, *run2* and best accuracies obtained among all participant systems are shown in Table 3.

	ECNUCS-team's system				
	spa-eng	*ita-eng*	*fra-eng*	*deu-eng*	average
run4	0.422	0.416	0.436	**0.452**	**0.432**
run3	0.408	0.426	**0.458**	0.432	0.431
	SOFTCARDINALITY-team's system				
	spa-eng	*ita-eng*	*fra-eng*	*deu-eng*	average
run1	**0.434**	**0.454**	0.416	0.414	0.430
run2	0.432	0.448	0.426	0.402	0.427

Table 3: Official results for our system and the top performing system ECNUCS (accuracies)

5 A Single Multilingual Model

This section presents the results of our additional experiments in search for a simplified model and in turn to respond to the following questions: i) Can one simplified-multilingual model overcome the approach presented in Section 4? ii) Does using additional circular-pivoting translations improve performance? and iii) Do other word-similarity functions work better than the q-grams measure?

First, it is important to note that the approach described in Section 4 used only patterns discovered in cardinalities. This means, that no language-dependent features was used, with the exception of the stemmers. Therefore, we wonder whether the patterns discovered in a pair of languages can be useful in other language pairs. To answer this question, a single prediction model was built by aggregating instances from each of the vector data sets into one data set with 4,000 training instances. Afterward, this model was used to provide predictions for the 2,000 test cases.

[1] https://translate.google.com

[2] http://snowball.tartarus.org

Moreover, customization for each pair of languages in the word-similarity function, which is show in Table 2, was set on the following unique set of parameters: $q_1 = 1$, $q_2 = 3$, $\alpha = 0.5$, $bias = 0.0$. Thus, the words are compared using q-grams and the Dice coefficient. In addition to the measure of q-grams, two "off-the-shelf" measures were used as nonparametric alternatives, namely: Edit Distance (Levenshtein, 1966) and the Jaro-Winkler similarity (Winkler, 1990).

In another attempt to simplify this model, we evaluated the predictive ability of each of the three groups of features shown in Table 1. The combination of groups 2 and 3, consistently obtained better results when the evaluation with 10 fold cross-validation was used in the training data. This result was consistent with the simple training versus test data evaluation. The sum of all previous simplifications significantly reduced the number of parameters and features in comparison with the model described in Section 4. That is, only one SVM and 4 parameters, namely: α, $bias$, q_1 and q_2.

Besides, the additional use of circular-pivoting translations was tested. In the original model, for every pair of texts (T_1, T_2) their pivot translations (T_1^t, T_2^t) were provided allowing the calculation of $|T_1 \cup T_2^t|$ and $|T_1^t \cup T_2|$. Translations T_1^t and T_2^t can also be translated back to their original languages obtaining T_1^{tt} and T_2^{tt}. These additional translations in turn allows the calculation of $|T_1^{tt} \cup T_2^t|$ and $|T_1^t \cup T_2^{tt}|$. This procedure can be repeated again to obtain T_1^{ttt} and T_2^{ttt}, which in turn provides $|T_1 \cup T_2^{ttt}|$, $|T_1^{ttt} \cup T_2|$, $|T_1^{tt} \cup T_2^{ttt}|$ and $|T_1^{ttt} \cup T_2^{tt}|$. The original feature set is denoted as *t*. The extended feature sets using double-pivoting translations and triple-pivot translations are denoted respectively as *tt* and *ttt*.

The results obtained with this simplified model using single, double and triple pivot translations are shown in Table 4. The first column indicates the word-similarity function used by the soft cardinality and the second column indicates the number of pivoting translations.

6 Discussion

In spite of the customization of the parameter C in the *run2*, the *run1* obtained better results than *run2* (see Table 3). This result indicates that the simpler model produced better predictions in unseen data.

Soft C.	#*t*	*spa-e*	*ita-e*	*fra-e*	*deu-e*	avg.
Ed.Dist.	*t*	0.444	0.450	0.440	0.410	0.436
Ed.Dist.	*tt*	0.452	0.464	0.434	0.432	0.446
Ed.Dist.	*ttt*	**0.464**	0.468	0.440	0.424	0.449
Jaro-W.	*t*	0.422	0.450	0.426	0.406	0.426
Jaro-W.	*tt*	0.430	0.456	0.444	0.400	0.433
Jaro-W.	*ttt*	0.426	0.458	0.430	0.430	0.436
q-grams	*t*	0.428	0.456	0.456	**0.432**	0.443
q-grams	*tt*	0.436	**0.478**	0.444	0.430	0.447
q-grams	*ttt*	0.452	0.474	**0.464**	0.442	**0.458**

Table 4: Single-multilingual model results (accuracies)

It is also important to note that two of the three multilingual systems proposed in Section 5 achieved higher scores than the best official results (see rows containing "*t*" in Table 4). This indicates that the proposed simplified model is able to discover patterns in the cardinalities of a pair of languages and project them into the other language pairs.

Regarding the use of additional circular-pivoting translations, Table 4 shows that *t* was overcome on average by *tt* and *tt* by *ttt* in all cases of the three sets of results. The relative improvement obtained by comparing *t* versus *ttt* for each group was 3.0% in Edit Distance, 2.3% for Jaro-Winkler and 3.4% for the q-gram measure. This same trend holds roughly for each language pair.

7 Conclusions

We described the SOFTCARDINALITY system that participated in the SemEval CLTE evaluation campaign in 2013, obtaining the best results in data sets *spa-eng* and *ita-eng,* and achieving the third place on average. This result was obtained using separate models for each language pair. It was also concluded that a single-multilingual model outperforms that approach. Besides, we found that the use of additional pivoting translations provide better results. Finally, the measure based on q-grams of characters, used within the soft cardinality, resulted to be the best option among other measures of word similarity. In conclusion, the soft cardinality method used in combination with SMT and SVM classifiers is a competitive method for the CLTE task.

Acknowledgments

This research was funded in part by the Systems and Industrial Engineering Department, the Office of Student Welfare of the National University of Colombia, Bogotá, and through a grant from the Colombian Department for Science, Technology and Innovation, Colciencias, proj. 1101-521-28465 with funding from "El Patrimonio Autónomo Fondo Nacional de Financiamiento para la Ciencia, la Tecnología y la Innovación, Francisco José de Caldas." The third author recognizes the support from Mexican Government (SNI, COFAA-IPN, SIP 20131702, CONACYT 50206-H) and CONACYT–DST India (proj. 122030 "Answer Validation through Textual Entailment").

References

Corinna Cortes and Vladimir N. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Kevin Duh, Ching-Man Au Yeung, Tomoharu Iwata, and Masaaki Nagata. 2013. Managing information disparity in multilingual document collections. *ACM Trans. Speech Lang. Process.*, 10(1):1:1–1:28, March.

Mark Hall, Frank Eibe, Geoffrey Holmes, and Bernhard Pfahringer. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Daniel S. Hirschberg. 1977. Algorithms for the longest common subsequence problem. *J. ACM*, 24(4):664–675, October.

M.A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, pages 414–420, June.

Sergio Jimenez, Fabio Gonzalez, and Alexander Gelbukh. 2010. Text comparison using soft cardinality. In Edgar Chavez and Stefano Lonardi, editors, *String Processing and Information Retrieval*, volume 6393 of *LNCS*, pages 297–302. Springer, Berlin, Heidelberg.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft cardinality+ ML: learning adaptive similarity functions for cross-lingual textual entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval, *SEM 2012)*, Montreal, Canada. ACL.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5):493–502, October.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24:377–439, December.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia*. Association for Computational Linguistics.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Detecting semantic equivalence and information disparity in cross-lingual documents. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, page 120–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 670–679, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. 2012. semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Canada.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, and Luisa Bentivogli. 2013. Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Gerard Salton, Andrew K. C. Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359. American Statistical Association.