SOFTCARDINALITY: Hierarchical Text Overlap for Student Response Analysis

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 used to participate in the Student-Response-Analysis task-7 at SemEval 2013. This system is based on text overlap through the soft cardinality and a new mechanism for weight propagation. Although there are several official performance measures, taking into account the overall accuracy throughout the two availabe data sets (Beetle and SciEntsBank), our system ranked first in the *2 way* classification task and second in the others. Furthermore, our system performs particularly well with "unseen-domains" instances, which was the more challenging test set. This paper also describes another system that integrates this method with the lexical-overlap baseline provided by the task organizers obtaining better results than the best official results. We concluded that the soft cardinality method is a very competitive baseline for the automatic evaluation of student responses.

1 Introduction

The Student-Response-Analysis (SRA) task consist in provide assessments of the correctness of student answers (A), considering their corresponding questions (Q) and reference answers (RA) (Dzikovska et al., 2012). SRA is the task-7 in the SemEval 2013 evaluation campaign (Dzikovska et al., 2013). The method used in our participation was basically text overlap based on the soft cardinality (Jimenez et al., 2010) plus a machine learning classifier. This method did not use any information external to the data sets except for a stemmer and a list of stop words.

The soft cardinality is a general model for object comparison that has been tested at text applications. Particularly, this text overlap approach has provided strong baselines for several applications, i.e. entity resolution (Jimenez et al., 2010), semantic textual similarity (Jimenez et al., 2012a), cross-lingual textual entailment (Jimenez et al., 2012b), information retrieval, textual entailment and paraphrase detection (Jimenez and Gelbukh, 2012). A brief description of the soft cardinality is presented in the next section.

The data for SRA consist of two data sets *Beetle* (5,199 instances) and *SciEntsBank* (10,804 instances) divided into training and test sets (76%-24% for *Beetle* and 46%-54% *SciEntsBank*). In addition, the test part of *Beetle* data set was divided into two test sets: "unseen answers" (35%) and "unseen questions" (65%). Similarity, *SciEntsBank* test part is divided into "unseen answers" (9%), "unseen questions" (13%) and "unseen domains" (78%). All texts are in English.

The challenge consists in predicting for each instance triple (Q, A, RA) an assessment of correctness for the student's answer. Three levels of detail are considered for this assessment: 2 way (*correct* and *incorrect*), 3 way (*correct*, *contradictory* and *incorrect*) and 5 way (*correct*, *incomplete*, *contradictory*, *irrelevant* and *non-in-the-domain*).

Section 3 presents the method used for the extraction of features from texts using the soft cardinality to provide a vector representation. In Section 4, the details of the system used to produce our predic-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 280–284, Atlanta, Georgia, June 14-15, 2013.

tions are presented. Besides, in that section a system that integrates our system with the lexical-overlap baseline proposed by the task organizers is also presented. This combined system was motivated by the observation that our system performed well in the *SciEntsBank* data set but poorly in *Beetle* in comparison with the lexical-overlap baseline. The results obtained by both systems are also presented in that section.

Finally in Section 5 the conclusions of our participation in this evaluation campaign are presented.

2 Soft Cardinality

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

$$|S|' = \sum_{i=1}^{n} w_i \cdot \left(\sum_{j=1}^{n} \mathbf{sim}(s_i, s_j)^p \right)^{-1} \quad (1)$$

Having $S = \{s_1, s_2, \ldots, s_n\}$; $w_i \geq 0$; $p \geq 0$; $1 > \mathbf{sim}(x, y) \geq 0$, $x \neq y$; and $\mathbf{sim}(x, x) = 1$. The parameter p controls the degree of "softness" of the cardinality (the larger the "harder"). In fact, when $p \to \infty$ the soft cardinality is equivalent to classical set cardinality. The default value for this parameter is $p = 1$. The coefficients w_i are weights associated with each element, which can represent the importance or informative character of each element. The function **sim** is a similarity function that compares pairs of elements in the collection S.

3 Features from Cardinalities

It is commonly accepted that it is possible to make a fair comparison of two objects if they are of the same nature. If the objects are instances of a compositional hierarchy, they should belong to the same class to be comparable. Clearly, a house is comparable with another house, a wall with another wall and a brick with another brick, but walls and bricks are not comparable (at least not directly). Similarly, in text applications documents should be compared with documents, sentences with sentences, words with words, and so on.

However, a comparison measure between a sentence and a document can be obtained with different approaches. First, using the information retrieval approach, the document is considered like a very long sentence and the comparison is then straight forward. Another approach is to make pairwise comparisons between the sentence and each sentence in the document. Then, the similarity scores of these comparisons can be aggregated in a single score using average, max or min functions. These approaches have issues, the former ignores the sentence subdivision of the document and the later ignores the similarities among the sentences in the document.

In the task at hand, each instance is composed of a question Q, a student answer A, which are sentences, and a collection of reference answers RA, which could be considered as a multi-sentence document. The soft cardinality can be used to provide values for $|Q|'$, $|A|'$, $|RA|'$, $|Q \cap A|'$, $|A \cap RA|'$ and $|Q \cap RA|'$. The intersections that involve RA require a special treatment to tackle the aforementioned issues.

Let's start defining a word-similarity function. Two words (or terms) t_1 and t_2 can be compared dividing them into character q-grams (Kukich, 1992). The representation in q-grams of t_i can be denoted as $t_i^{[q]}$. Similarly, a combined representation using a range of q-grams of different length can be denoted as $t_i^{[q_1:q_2]}$. For instance, if t_1 ="home" then $t_1^{[2:3]}$={"ho","om","me","hom","ome"}. Thus, $t_1^{[q_1:q_2]}$ and $t_2^{[q_1:q_2]}$ representations can be compared using the Dice's coefficient to build a word-similarity function:

$$\mathbf{sim_{words}}(t_1, t_2) = \frac{2 \cdot \left| t_1^{[q_1:q_2]} \cap t_2^{[q_1:q_2]} \right|}{\left| t_1^{[q_1:q_2]} \right| + \left| t_1^{[q_1:q_2]} \right|} \quad (2)$$

Note that in eq. 2 the classical set cardinality was used, i.e $|x|$ means classical cardinality and $|x|'$ soft cardinality.

The function $\mathbf{sim_{words}}$ can be plugged in eq.1 to obtain the soft cardinality of a sentence S (using unitary weights $w_i = 1$ and $p = 1$):

$$|S|' = \sum_{i=1}^{|S|} \left(\sum_{j=1}^{|S|} \mathbf{sim_{word}}(t_i, t_j) \right)^{-1} \quad (3)$$

$\lvert X\rvert$	$\lvert Y\rvert$	$\lvert X \cup Y\rvert$
BF1: $\lvert Q\rvert'$	BF2: $\lvert A\rvert'$	BF3: $\lvert Q \cup A\rvert'$
BF2: $\lvert A\rvert'$	BF4: $\lvert RA\rvert''$	BF5: $\lvert RA \cup A\rvert''$
BF1: $\lvert Q\rvert'$	BF4: $\lvert RA\rvert''$	BF6: $\lvert RA \cup Q\rvert''$

Table 1: Basic feature set

Where t_i are the words in the sentence S .

The sentence-soft-cardinality function can be used to build a sentence-similarity function to compare two sentences S_1 and S_2 using again the Dice's coefficient:

$$\mathbf{sim}_{\mathbf{sent.}}(S_1, S_2) = \frac{2 \cdot (|S_1|' + |S_2|' - |S_1 \cup S_2|')}{|S_1| + |S_2|} \quad (4)$$

In this formulation $S_1 \cup S_2$ is the concatenation of both sentences.

The eq. 4 can be plugged again into eq. 1 to obtain the soft cardinality of a "document" RA, which is a collection of sentences $RA = \{S_1, S_2 \ldots, S_{|RA|}\}$:

$$|RA|'' = \sum_{i=1}^{|RA|} |S_i|' \cdot \left(\sum_{j=1}^{|RA|} \mathbf{sim}(S_i, S_j) \right)^{-1} \quad (5)$$

Note that the soft cardinalities of the sentences $|S_i|'$ were re-used as importance weights w_i in eq. 1. These weights are propagations of the unitary weights assigned to the words, which in turn were aggregated by the soft cardinality at sentence level (eq. 3). This soft cardinality is denoted with double apostrophe because is a function recursively based in the single-apostrophized soft cardinality.

The proposed soft cardinality expressions are used to obtain the basic feature set presented in Table 1. The soft cardinalities of $|Q|'$, $|A|'$ and $|Q \cup A|'$ are calculated with eq. 3. The soft cardinalities $|RA|''$, $|RA \cup A|''$ and $|RA \cup Q|''$ are calculated with eq. 5. Recall that $Q \cup A$ is the concatenation of the question and answer sentences. Similarly, $RA \cup A$ and $RA \cup Q$ are the collection of reference answers adding A xor Q .

Starting from the basic feature set, an extended set, showed in Table 2, can be obtained from each one of the three rows in Table 1. Recall that $|X \cap Y| = |X| + |Y| - |X \cup Y|$ and $|X \setminus Y| = |X| - |X \cap Y|$.

EF1:	$\lvert X \cap Y\rvert$	EF2:	$\lvert X \setminus Y\rvert$
EF3:	$\lvert Y \setminus X\rvert$	EF4:	$\frac{\lvert X \cap Y\rvert}{\lvert X\rvert}$
EF5:	$\frac{\lvert X \cap Y\rvert}{\lvert Y\rvert}$	EF6:	$\frac{\lvert X \cap Y\rvert}{\lvert X \cup Y\rvert}$
EF7:	$\frac{2 \cdot \lvert X \cap Y\rvert}{\lvert X\rvert + \lvert Y\rvert}$	EF8:	$\frac{\lvert X \cap Y\rvert}{\sqrt{\lvert X\rvert \cdot \lvert Y\rvert}}$
EF9:	$\frac{\lvert X \cap Y\rvert}{\min(\lvert X\rvert, \lvert Y\rvert)}$	EF10:	$\frac{\lvert X \cap Y\rvert}{\max(\lvert X\rvert, \lvert Y\rvert)}$
EF11:	$\frac{\lvert X \cap Y\rvert \cdot (\lvert X\rvert + \lvert Y\rvert)}{2 \cdot \lvert X\rvert \cdot \lvert Y\rvert}$	EF12:	$\lvert X \cup Y\rvert - \lvert X \cap Y\rvert$

Table 2: Extended feature set

$Y|$. Consequently, the total number of features is 6 basic features plus 12 extended features multiplied by 3, i.e. 42 features. 42 = 12*3 + 6

4 Systems Description

4.1 Submitted System

First, each text in the SRA data was preprocessed by tokenizing, lowercasing, stop-words[1] removing and stemming with the Porter's algorithm (Porter, 1980). Second, each stemmed word t was represented in q-grams: $t^{[3:4]}$ for *Beetle* and $t^{[4]}$ for *SciEntsBank*. These representations obtained the best accuracies in the training data sets.

Q 问题
A 答案
RA 参考答案

Two vector data sets were obtained extracting the 42 features–described in Section 3–for each instance in *Beetle* and *SciEntsBank* separately. Then, three classification models (*2 way*, *3way* and *5 way*) were learned from the training partitions on each vector data set using a J48 graft tree (Webb, 1999). All 6 resulting classification models were boosted with 15 iterations of bagging (Breiman, 1996). The used implementation of this classifier was that included in WEKA v.3.6.9 (Hall et al., 2009). The results obtained by this system are shown in Table 3 in the rows labeled with "Soft Cardinality-run1".

每个Q-A-RA实例得到42个特征

用j48 决策树分类，在weka中获得

weka是免费的数据挖掘工具包

4.2 An Improved System

At the time when the official results were released, we observed that our submitted system performed pretty well in *SciEntsBank* but poorly in *Beetle*. Moreover, the lexical-overlap baseline outperformed our system in *Beetle*. Firstly, we decided to include in our feature set the 8 features of the lexical overlap baseline described by Dzikovska et al. (2012)

[1] those provided by `nltk.org`

soft cardinality + lexical overlap baseline

		Beetle			SciEntsBank					
Task	System	UA[1]	UQ[2]	All	UA[1]	UQ[2]	UD[3]	All	All	Rank
2 way	*Soft Cardinality-unofficial*	*0.797*	*0.725*	*0.750*	*0.717*	*0.733*	*0.726*	*0.726*	***0.730***	-
	Soft Cardinality-run1	0.781	0.667	0.707	0.724	0.745	0.711	0.716	0.715	1
	ETS-run1	0.811	0.741	0.765	0.722	0.711	0.698	0.702	0.713	2
	CU-run1	0.786	0.718	0.742	0.656	0.674	0.693	0.687	0.697	3
	Lexical overlap baseline	0.797	0.740	0.760	0.661	0.674	0.676	0.674	0.690	6
3 way	*Soft Cardinality-unofficial*	*0.608*	*0.532*	*0.559*	*0.656*	*0.671*	*0.646*	*0.650*	***0.634***	-
	ETS-run1	0.633	0.551	0.580	0.626	0.663	0.632	0.635	0.625	1
	Soft Cardinality-run1	0.624	0.453	0.513	0.659	0.652	0.637	0.641	0.618	2
	CoMeT-run1	0.731	0.518	0.592	0.713	0.546	0.579	0.587	0.588	3
	Lexical overlap baseline	0.595	0.512	0.541	0.556	0.540	0.577	0.570	0.565	8
5way	*Soft Cardinality-unofficial*	*0.572*	*0.476*	*0.510*	*0.552*	*0.520*	*0.534*	*0.534*	***0.530***	-
	ETS-run1	0.574	0.560	0.565	0.543	0.532	0.501	0.509	0.519	1
	Soft Cardinality-run1	0.576	0.451	0.495	0.544	0.525	0.512	0.517	0.513	2
	ETS-run2	0.715	0.621	0.654	0.631	0.401	0.476	0.481	0.512	3
	Lexical overlap baseline	0.519	0.480	0.494	0.437	0.413	0.415	0.417	0.430	11
Total number of test instances		439	819	1,258	540	733	4,562	5,835	7,093	

TEST SETS: unseen answers[1], unseen questions[2], unseen domains[3].

Table 3: Official results for the top-3 performing systems (among 15), the lexical overlap baseline in the SRA task SemEval 2013 and unofficial results of the soft cardinality system combined with the lexical overlap (in italics). Performance measure used: overall accuracy.

(see Text::Similarity::Overlaps[2] package for more details).

Secondly, the lexical overlap baseline aggregates the pairwise scores between each reference answer and the student answer by taking the maximum value of the pairwise scores. So, we decided to use this aggregation mechanism instead of the aggregation proposed through eq. 3.

Thirdly, only at that time we realized that, unlike *Beetle*, in *SciEntsBank* all instances have only one reference answer. Consequently, the only effect of eq. 5 in *SciEntsBank* was in the calculation of $|RA \cup A|''$ (and $|RA \cup Q|''$) by $|X \cup Y|'' = \frac{|X|'+|Y|'}{1+\mathbf{sim_{sent.}}(X,Y)}$. As a result, this transformation induced a boosting effect in $X \cap Y$ making $|X \cap Y|'' \geq |X \cap Y|'$ for any X, Y. We decided to use this intersection-boosting effect not only in $RA \cap A$, $RA \cap Q$, but in $Q \cap A$. This intersecton boosting effect works similarly to the Lesk's measure (Lesk, 1986) included in the lexical overlap baseline.

交叉增强

The individual effect in the performance of each of the previous decisions was positive in all three cases. The results obtained using an improved system that implemented those three decisions are shown in Table 3–in italics. This system would have obtained the best general overall accuracy in the official ranking.

5 Conclusions

We participated in the Student-Response-Analysis task-7 in SemEval 2013 with a text overlap system based on the soft cardinality. This system obtained places 1st (*2 way* task) and 2nd (*3 way* and *5 way*) considering the overall accuracy across all data sets and test sets. Particularly, our system was the best in the largest and more challenging test set, namely "unseen domains". Moreover, we integrated the lexical overlap baseline to our system obtaining even better results.

As a conclusion, the text overlap method based on the soft cardinality is very challenging base line for the SRA task.

[2] http://search.cpan.org/dist/Text-Similarity/lib/Text/Similarity/Overlaps.pm

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

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: a dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, page 200–210, Stroudsburg, PA, USA. Association for Computational Linguistics.

Myroslava O. Dzikovska, Rodney D. Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (*SEM 2013)*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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

Sergio Jimenez and Alexander Gelbukh. 2012. Baselines for natural language processing tasks. *Appl. Comput. Math.*, 11(2):180–199.

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. 2012a. Soft cardinality: A parameterized similarity function for text comparison. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval, *SEM 2012)*, Montreal, Canada.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012b. 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 Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24:377–439, December.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, page 24–26, New York, NY, USA. ACM.

Martin Porter. 1980. An algorithm for suffix stripping. *Program*, 3(14):130–137, October.

Geoffrey I. Webb. 1999. Decision tree grafting from the all-tests-but-one partition. In *Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2*, IJCAI'99, pages 702–707, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.