Sentence Difficulty Analysis with Local Feature Space and Global Distributional Difference

Young-Bum Kim[1], YoungJo Kim[2], and Yu-Seop Kim[3,4,*]

[1] Dept. of Computer Science, University of Wisconsin-Madison,
1210 W. Dayton St. Madison, WI 53706-1685
stylebbum@gmail.com
[2] Dept. of Computer Science, Brown University, Providence, Rhode Island 02912
Youngjo_kim@brown.edu
[3] Dept. of Ubiquitous Computing, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702, Korea
[4]Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea
yskim01@hallym.ac.kr

Abstract. In this paper, we consider the problem of sentence difficulty analysis from various angles. Past works have endeavored to design deterministic scoring algorithms depending only on semantic and syntactic information. We propose instead not only to hire local feature space representing individual sentence with its syntactic and semantic structure, but also to consider global distributional difference among corpora. For the local feature space, we select 28 linguistic features and transform them into conjuncted and discretized form. By applying global score classification, we can show its much improved results. We test our proposed model to 1,000 sentences and get much higher accuracy than traditional learning models such as SVM and AdaBoost.

Keywords: Sentence Difficulty, Local Feature Space, Global Distributional Difference, Syntactic Structure, Semantic Structure, SVM, AdaBoost.

1 Introduction

The estimation of sentence complexity and difficulty has played a role in various research areas, such as psychology, psycholinguistics, neurolinguistics and education, from the use of reading index, mean length of utterance, and related scores in child language development[1,2], to complexity scores related to reading difficulty in human sentence processing studies[3,4]. Of course, without manual annotation and deterministic scoring algorithms, scoring of such sentence difficulty is infeasible. However, autonomous estimation of such measures with natural language processing and machine learning techniques definitely can have high utility in terms of reduction of time which is required to annotate and score samples.

The sentence difficulty estimation could be originated from the study of the Reading Index[1]. Reading Index simply assumes that the longer the sentence is, the more

* Corresponding author.

G. Lee et al. (Eds.): ICHIT 2012, LNCS 7425, pp. 716–722, 2012.

difficult to understand it is. This is too simple to be adapted to the sentence difficulty estimation. [5] proposed the word maturity to estimate the single word difficulty level. It is also used for the sentence difficulty estimation by combining the individual word difficulty into the sentence difficulty. However, this approach could not consider the syntactic complexity of the sentence. The syntactic and semantic surprisal[6, 7] tried to consider the syntactic and semantic structure concurrently and to estimate the sentence difficulty. Although this approach tries to consider both aspects, the semantic surprisal has been shown to be more related to the syntactic difficulty.

In this paper, we propose a new approach not only to hire local feature space representing individual sentence with its syntactic and semantic structure, but also to consider global distributional difference among corpora. Each sentence includes semantic and syntactic aspects, individually tied with various linguistic features transformed by conjunction and discretization. After the local learning with above feature, we develop another novel global score classification method using the distributional difference between training and test corpus. When deciding the difficulty level more precisely, we hire the well-known distribution of training corpus. We apply our model to 1,000 sentences, and induce semantic and syntactic difficulty with substantially higher accuracy than the result from traditional learning models such as SVM and AdaBoost. Our analysis can also explain various components which human also considers as an important one deciding the difficulty.

We introduce how to build the local feature space used to decide local difficulty level in section 2. Section 3 explains our model integrating the local and the global score. We show our experimental results in section 4, and discuss the conclusion and future works in section 5.

2 Feature Generation

Local feature space is closely related to the sentence-level difficulty. It is also considered to be related to syntactic and semantic component of the sentence. To build the space, we firstly collect raw linguistic sentential features seemed to affect the difficulty. We collect 28 features and table 1 shows five features and their description for example. We also include syntactic and semantic surprisal proposed by [6,7] in the feature set.

We apply discretization to build a specific feature set affecting sentential and both components, syntactic and semantic, difficulty level. In this research, discretization of continuous features using minimal entropy [8,9] is used. This algorism uses the information entropy of each class to find optimal bin partition. Following the notation of Fayyad and Irani, with a given set of instances S, a feature A, and a cut value T, we denote the information entropy $L'(A,T;S)$ as:

$$L'(A,T;S) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2), \quad (1)$$

where S_1 is a subset of S with values of A which is not greater than T and $S_2 = S - S_1$.

Table 1. Raw Features

Features	Description
Depth	Depth of parsed tree
Length	Number of alphabets in sentence
Similarity	Similarity of translated sentences
Proper Nouns	Number of proper nouns
Verb Count	Number of verbs in sentence

The value T_A which minimizes the entropy is the cut point of discretization. Fayyad and Irani set stop condition with MDLP (Minimal Description Length Principle) for this recursive discretization given by:

$$Entropy(S) - L'(A,T;S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}, \quad (2)$$

where N is the number of instances in the set S,

$$\Delta(A,T;S) = \log_2(3^k - 2) \quad -[k \cdot Entropy(S) - k_1 \cdot Entropy(S_1) - k_2 \cdot Entropy(S_2)], \quad (3)$$

where k_i is the number of class labels in the set S_i.

We also generate new features by making its conjunctions. At first, we make conjunction of every two features, ${}_nC_2$, where n is the number of raw features. Like the raw features, these new features are also discretized.

3 Our Model

We initially analyze each sentence according to the prediction from a local learning stage. As the next stage reanalyzes the initial prediction results, the local procedure seeks to minimize training error to take a highly over-fitted set which can show a definite local trend. For each sentence, $s \in S$, where S is the test corpus, we consider all possible semantic and syntactic difficulty pair of s. Consequently, we compute the following scoring function. We classify every score into syntactic, semantic and sentential score. The sentential score is acquired by adding another two scores.

$$TOTAL\_SCORE_i(s,S) = \alpha * LOCAL\_SCORE_i(s) + \beta * GLOBAL\_SCORE_i(S), \quad (4)$$

where α and β are the parameters. *i* is difficulty level from 1 to 5. It means that one sentence can have 5 difficulty level values and the level with the biggest value is decided to be the level of the sentence. The Local_Score() is given to each sentence in view of its syntactic and semantic difficulty, whereas the Global_Score() is given to each test corpus including the sentence which is now investigated.

The leaning procedure is described in table 2.

Table 2. The Learning Procedure

Let θ is the set of initial label in corpus S, where each component of θ is the label of each sentence.

While (!convergence)

θ = shuffle(θ)

For i = 0 to n where n is the number of test sentences

$\theta' = \theta - \{\theta_i\}$, where θ_i is the label of i^{th} sentence s_i.

$\theta_{new} = \theta' \cup \arg\max_{j \in L} TOTAL\_SCORE_j(s_i, S - \{s_i\})$

, where $1 \leq j \leq$ L which L is the number of labels.

Local_Score, and also Global_Score, is composed of three scores, syntactic, semantic, and sentence score. These scores are computed separately and consequently added up to the Local and Global score. For the given sentence *s*, a feature vector ***x*** is constructed with discretization for feature selection.

*Local_Score$_i$(**x**)*, for label *i*, is computed as follows.

$$Local\_Score_i(\vec{x}) = \frac{Score_i(\vec{x})}{\sum_{1 \leq j \leq L} Score_j(\vec{x})} \tag{5}$$

And *Score$_i$(**x**)* for label *i*, is calculated as following.

$$Score_i(\vec{x}) = \sum_{1 \leq j \leq F} (w_{i,j} * f_j), \tag{6}$$

where w_{ij} is weight for the connection between a feature component, f_j, and a label, l_i. F is the number of the elements in vector ***x***.

Global_Score, representing the distributional distance between the training corpus and the test corpus is also computed. This score of corpus S is also composed of three components and is computed as follows.

$$Global\_Score_i(S) = \frac{1}{D_{KL}(S \parallel S^{train}) + D_{KL}(S^{train} \parallel S)}, \tag{7}$$

where $D_{KL}(a||b)$ means the Kullback-Leibler divergence [10] between *a* and *b*. *S* is the corpus now including the test sentence and S^{train} is the corpus used in the training phase.

4 Experimental Results

In this section we will show the experimental environment and its results.

4.1 Corpus

To test with our hierarchical feature space, we apply it to a set of one thousand English sentences extracted from Korea Herald[1] which is an English newspaper publishing in Korea. The set includes approximately 20,000 English words. Each sentence is manually graded separately in view of its syntactic and semantic difficulty. The grade could be from level 1 (easiest) to level 5 (hardest). Table 3 shows the level distribution of the level-annotated corpus

Table 3. Corpus Statistics for the 1,000 Sentences

Labels	Distribution
Syntactic Level	
1	0.079
2	0.283
3	0.362
4	0.267
5	0.009
Semantic Level	
1	0.091
2	0.447
3	0.367
4	0.093
5	0.002

4.2 Feature Selection

We firstly build 28 raw linguistic features representing a sentence. From these, we select features more affecting the sentential difficulty in view of its syntax and semantics. We use the discretization method described in section 2. After discretization, 10 and 5 features are selected as the syntactic and semantic features, respectively. The number of syntactic features selected is double of the semantic ones. It caused by the fact that many sentential features are more directly representing syntactic characteristics of the sentence because the syntax is more easily to be described.

We also conjunct two features and discretize them. We generate ${}_{28}C_2 = 378$ features. After discretization, we can get 67 syntactic features and only 12 semantic features. The difference of the number is supposed to be made by the fact that the semantic difficulty is more complicated to be explained with the raw features.

[1] http:// http://www.koreaherald.com/

Table 4. Experimental Results. (1) is surprisals used as a base line. (2) uses 28 Raw Features and (3) discretizes original raw features. (4) discretizes and conjuncts features and (5) discretizes raw features and conjuncted features.

	Multiclass SVM		Adaboost		Our Model	
	Syn	Sem	Syn	Sem	Syn	Sem
(1)	26%	35%	31%	39%	59%	51%
(2)	39%	38%	45%	49%	62%	56%
(3)	65%	62%	59%	57%	71%	68%
(4)	72%	68%	61%	55%	75%	78%
(5)	75%	71%	69%	62%	82%	86%

4.3 Feature Selection

As our baseline, we use syntactic and semantic Surprisal values of sentences extracted from the wall street journal corpus as a knowledge space, which are integrated measures to analyze the reading time [7]. Before applying our model, we first treat each of one thousand sentences in turn as the test sentence, with the other 999 serving as training examples to get a "good" starting point. For each test instance, we iterate the replacement procedure until convergence. The number of required iterations varies from 2 to 15 (depending on the start instance and constant values), and each iteration takes no more than 4 seconds of run-time on a 3.4GHz Intel i7 Processor.

Table 4 shows the experimental results. The baseline without any feature selection shows only 26% and 35% for the syntactic and semantic difficulty estimation. However, our model with all kinds of feature selection shows much increased accuracy, up to 82% and 86% for its syntactic and semantic difficulty estimation.

5 Concluding Remarks

We estimate the sentence difficulty level with two aspects. First, we try to estimate the sentence difficulty itself with syntactic and semantic views. With selected features conjuncted and discretized from the raw ones, the sentence difficulty is computed. Second, we add the corpus-level information to the initial estimated difficulty. This local and global estimator increases the accuracy up to 24% points.

For the future work, we will implement an application using this difficulty level estimation. This research is originally motivated from e-learning application. By estimating the sentence level automatically, the second language learners can be given more proper English sentences after their assessment. And more semantic feature should be mined. Many of features proposed in this research are basically syntax-based features.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation(NRF) funded by the Ministry of Education, Science and Technology(2010-0010612).

References

1. Bormuth, J.R.: Readability: A New Approach. Reading Research Quarterly 1(3), 79–132 (1966)
2. Klee, T., Fitzgerald, D.: The Relation between Grammatical Development and Mean Length of Utterance in Morphemes. Journal of Child Language 12, 251–269 (1985)
3. Taylor, W.L.: Cloze Procedure: A New Tool for Measuring Readability. Journalism Quarterly 30, 415–433 (1953)
4. Dubay, W.H.: The Principles of Readability. Impact Information, Costa Mesa (2004)
5. Kireyev, K., Landauer, T.K.: Word Maturity: Computational Modeling of Word Knowledge. In: 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT (2011)
6. Roark, B., Bachrach, A., Cardenas, C., Pallier, C.: Deriving Lexical and Syntactic Expectation-based Measures for Psycholinguistic Modeling via Incremental Top-down Parsing. In: 2009 Conference on Empirical Methods in Natural Language Processing, pp. 324–333 (2009)
7. Mitchell, J., Lapata, M., Demberg, V., Keller, F.: Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure. In: 48th Annual Meeting of the Association for Computational Linguistics, pp. 196–206 (2010)
8. Catlett, J.: On Changing Continuous Attributes into Ordered Discrete Attributes. In: Kodratoff, Y. (ed.) EWSL 1991. LNCS, vol. 482, pp. 164–178. Springer, Heidelberg (1991)
9. Fayyad, U.M., Irani, K.B.: Multi-Interval Discretization of Continuous-valued Attributes for Classification Leaning. In: International Joint Conference on Artificial Intelligence, pp. 1022–1027 (1993)
10. Kullback, S., Leibler, R.A.: On Information and Sufficiency. Annals of Mathematical Statistics 22(1), 79–86 (1951)