The Use of Metrics for Measuring Informality Levels in Web 2.0 Texts

Alejandro Mosquera[1], **Paloma Moreda**[1]

[1] Department of Language and Computing Systems – University of Alicante
Alicante – Spain

{amosquera,moreda}@dlsi.ua.es

Abstract. *The study of text informality can provide us with valuable information for different NLP tasks. In the particular case of social media texts, their special characteristics like the presence of emoticons, slang or colloquial words can be used for obtaining additional information about their informality level. This paper demonstrates that the discovery of informality levels in Web 2.0 texts can be improved by incorporating formality and informality scores. The classification method based on our proposal reaches a 78% F1 using unsupervised machine learning techniques.*

1. Introduction

As the Web increases its importance and popularity new studies appear about the Internet-specific language. With the evolution of Web 2.0, we can differentiate a new variety of text types like blog posts, *tweets* or chat conversations. The absence of subordinate constructions, presence of slang, *netspeak*, chat-style abbreviations, emoticons and colloquial expressions are just some characteristics of their language [Squires 2010].

The informal nature of these new text types and their characteristics represents a challenge for the existing Natural Language Processing (NLP) applications. The first step to approach this challenge is being able to objectively quantify their informality level. This would yield valuable information for NLP tasks such as sentiment analysis, information extraction or machine reading.

For this reason, in this paper we are going to propose a new method for improving the task of discovering informality levels in Web 2.0 texts based on formality and informality scores. To do this, we also propose a new metric based on text characteristics, the I-Measure, used along this study with unsupervised machine learning techniques.

This paper is organized as follows: In Section 2 we review the state of the art. Section 3 introduces the used metrics and describes our methodology. In Section 4, the experimental results of our method are analized. Finally, our main conclusions and future works are drawn in Section 5.

2. Related Work

The use of formality scores usually involves formulae based on text features. One of the more remarkable scores is the F-Measure [Heylighen and Dewaele 1999], that linguistically differentiates between two word groups, deictic and non deictic. The first group, pronouns, verbs, adverbs and interjections, increment their frequency on informal texts, otherwise the second group, nouns, adjectives and prepositions lower their frequency on

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 184–188,
Cuiabá, MT, Brazil, October 24–26, 2011.

informal texts. This score based on relationship between part-of-speech (POS) tags was used for characterizing the sentence-level formality of texts from Web sources and its distribution across different Internet text types [Lahiri et al. 2011].

In this study we propose a method for discovering informality levels having in mind the special characteristics of Web 2.0 texts. In order to extend the information obtained with the existing formality score, an informality measure will be developed and used with unsupervised machine learning techniques.

3. Measuring Informality

With the main objective of obtaining informality levels in social media texts, we hypothesise that those levels can be inferred by grouping texts taking into account their informality score. Hereby a score-based approach have been used in this work using the Expectation Maximization (EM) [Dempster et al. 1977] unsupervised machine learning algorithm.

3.1. Text Characteristics

A set of 22 characteristics was defined in order to obtain information about the formality and informality levels. Simplicity was the main criteria for our characteristic election, to minimize the possible errors introduced by NLP tools.

The part-of-speech tagger TreeTagger [Schmid 1994] was used for obtain all POS characteristics like the frequency of verbs, adverbs, prepositions, interjections, adjectives or nouns. We also include characteristics relative to sentence and word length like the average word and sentence length, using the POS information to determine the end of sentences.

We relied on regular expressions and heuristic rules for the English language in order to discover emoticons and wrong-typed words. A spell checker proved impractical besides computationally expensive. For this reason we used a small set of heuristic rules to detect common case typos e.g ”After the end of any sentence the next word must start with upper-case”.

Additionally, we detected unknown, slang, informal and offensive words by checking the presence of the lemma or the complete word in on-line dictionaries and parsing the obtained query results. We chose Wiktionary [Wikimedia Foundation 2011], Online Slang Dictionary [The Online Slang Dictionary 2011] and Advanced Learner Cambridge Online Dictionary [Cambridge University Press 2011] because these dictionaries can provide special description tags (Informal, Colloquial, Onomatopoeia, Offensive, Slang, Internet slang and Internet).

3.2. Formality/Informality Measures

The formality measure F-Measure was defined by Heylighen as follows:

F-Measure = (noun frequency + adjective freq. + preposition freq. + article freq. - pronoun freq. - verb freq. - adverb freq. - interjection freq. + 100)/2

Using a formality score based on POS tags can give us information about text informality, but in order to obtain a specific metric adapted to the Web 2.0 texts we have

to rely on its informal nature rather than grammatical information only. The characterization of special Web 2.0 text characteristics like non-standard abbreviations, colloquial expressions or presence of slang words give us additional information for discovering informality levels.

The most relevant features for our informality score were explored by a statistical method of factor analysis [Rummel 1970]. The varimax rotation criterium, that searches the rotated loadings that maximize the variance of the squared loadings for each factor, is one of the common mathematical procedures for accomplish factor analysis and was used to obtain the rotated matrix of factor loadings.

In a two-factor model with all loadings less than 0.4 suppressed for being considered not relevant, the frequency of wrong-typed words and the frequency of interjections obtained the higher loadings in factor 1 and 2 respectively.

As factors can not be measured directly a variable reduction technique, Principal Component Analysis (PCA) [Jolliffe 2002] was used. Using this technique we reduce the number of observed variables to a smaller number of uncorrelated principal components which represent most of the variance of the observed variables.

After obtaining three non-correlated variables with PCA analysis we define the I-Measure score as follows:

*I-Measure = (Wrong-typed Words freq. + Interjections freq. + Emoticon freq.) * 100*

3.3. Classification Algorithm

A two-cluster classification is used to identify less informal and more informal texts. In our proposal, the Expectation-Maximization (EM) clustering algorithm was used in one and two dimensions with the I-Measure and the F-Measure as unique features.

4. Evaluation and Results

The usual measures for performance evaluation in text classification algorithms are precision and recall. But in text clustering, the evaluation measurements cannot be taken directly, as we are mapping classes to clusters and each obtained cluster will have their own values. For this reason, the weighted average of the precision and recall of each cluster [Andritsos et al. 2003] was computed with the F1 score as:

$$P = \sum_{i=1}^{k} \frac{|Gi|}{|T|} Pi \qquad R = \sum_{i=1}^{k} \frac{|Gi|}{|T|} Ri \qquad F1 = 2\frac{PR}{P+R}$$

Where **G** is the number of texts assigned to the current class, **T** is the total number of texts and **k** is the number of clusters mapped to informality levels.

4.1. Corpus Characteristics

All our tests have been done with a subset of the Fundacion Barcelona Media corpus [Fundacion Barcelona Media 2009] that includes texts from the following Web 2.0 sources: **Slashdot**, a technology-related news website; **Ciao**, an online-shopping and product review portal; **Kongregate**, an on-line gaming and chat website; **Twitter**, a social networking and microblogging service; **MySpace**, a social networking website; **Digg**, a news voting and review website; and **Engadget**, and electronic products review portal.

Algorithm	Small Corpus			Big Corpus		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline	0.691	0.657	0.683	-	-	-
EM I-Measure	0.747	0.697	0.721	0.775	0.739	0.757
EM F-Measure	0.668	0.645	0.656	0.654	0.652	0.653
EM F-Measure & I-Measure	0.756	0.725	**0.740**	0.795	0.773	**0.784**

Table 1. Experimental results with small (350 texts) and big corpora (700 texts).

For evaluating the results, two volunteers have annotated the corpus texts by hand in two categories: "neutral" or "informal" regarding their informality level. To avoid the possibility that the classification may occur just by chance we used the Cohen's Kappa value [Cohen 1960], defined as $K = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$ (Where *Pr(a)* is the relative observed agreement among the evaluators and *Pr(e)* is the probability of agreement by chance) obtaining a 0.815 K that can be considered a good value.

4.2. Baseline

In our previous studies [Mosquera and Moreda 2011] a binary classification was developed. Using the K-Means clustering algorithm [Hartigan and Wong 1979] with a set of features extracted from a 350 text corpus (Mean sentence length, Freq. Non-printable words, Freq. prepositions, Freq. Nouns, Freq. Emoticons, Freq. Upper-case words and Freq. Informal words) we obtained a 68% F1.

4.3. Results

For evaluating with the same test environment that in our baseline, we split the corpus in two regarding their size, small with 350 texts and big with 700 texts. With the small corpus, the clustering of the I-Measure and F-Measure features with the EM algorithm scored the best F1 (74%), enhancing the results obtained in our baseline by a 8.82%, (see Table 1).

Using the big corpus the results showed a 10% improvement respect our baseline (78% F1) using the F-Measure and the I-Measure in a two-dimensional cluster (see Table 1), concluding that the use of more text instances helps the clustering process to describe the model with more accuracy.

5. Conclusions and Future Works

In this study, we proposed a method for discovering two levels of informality in Web 2.0 texts. In order to achieve this, an informality metric based on different text characteristics was developed and used in combination with another formality measure. In addition, we have experimented with a number of different classification algorithms, obtaining a 78% F1 using the EM unsupervised machine learning algorithm.

Although the presented two-level classification shows the need of more clusters for a better understanding of the social media text types, the proposed combination of scores provides a valuable source of information about the text informality, being our method scalable in more text features and informality levels.

The future directions for expanding this work include more informality levels, the exploration of another classification algorithms and the addition of more text features.

Acknowledgments

This paper has been partially supported by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educación - Generalitat Valenciana (grant no. PROMETEO/2009/119, ACOMP/2010/286 and ACOMP/2011/001)

References

Andritsos, P., Tsaparas, P., Miller, R. J., and Sevcik, K. C. (2003). Limbo: A scalable algorithm to cluster categorical data. Technical report, University of Toronto, Department of Computer Science.

Cambridge University Press, C. U. P. (2011). Cambridge dictionary of american english. http://dictionary.cambridge.org/dictionary/british/.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Dempster, A. P., Laird, M. N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22.

Fundacion Barcelona Media, F. B. M. (2009). Caw 2.0 training datasets available from http://caw2.barcelonamedia.org.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Heylighen, F. and Dewaele, J.-M. (1999). Formality of language: definition, measurement and behavioral determinants. Technical report, Free University of Brussels.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, second edition.

Lahiri, S., Mitra, P., and Lu, X. (2011). Informality judgment at sentence level and experiments with formality score. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pages 446–457. Springer-Verlag.

Mosquera, A. and Moreda, P. (2011). Caracterización de niveles de informalidad en textos de la web 2.0. *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) (pending publication)*, 47.

Rummel, R. (1970). *Applied Factor Analysis*. Evanston: Northwestern University Press.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Squires, L. (2010). Enregistering internet language. *Language in Society*, 39(04):457–492.

The Online Slang Dictionary, T. O. S. D. (2011). http://onlineslangdictionary.com/.

Wikimedia Foundation, W. F. (2011). Wiktionary: The free dictionary. http://en.wiktionary.org/.