Testing a Statistical Word Stemmer based on Affixality Measurements in INEX 2012 Tweet Contextualization Track

Carlos-Francisco Méndez-Cruz[1], Edmundo-Pavel Soriano-Morales[1], and Alfonso Medina-Urrea[2]

[1] GIL-Instituto de Ingeniería UNAM, México
cmendezc@ii.unam.mx, sorianopavel@gmail.com
[2] El Colegio de México A.C., México
amedinau@colmex.mx

Abstract. This paper presents an experiment of statistical word stemming based on affixality measurements. These measurements quantify three characteristics of language. In this experiment we tested one strategy of stemming with three different sizes of training data. The developed stemmer was used by the automatic summarization system CORTEX to preprocess input texts and produce readable summaries. All summaries were evaluated as part of the INEX 2012 Tweet Contextualization Track. We present the results of evaluation and a discussion about our stemming strategy.

Key words: INEX, Automatic summarization system, Affixality Measurements, Morphological Segmentation, Statistical Stemming, CORTEX, Tweet Contextualization.

1 Introduction

The task proposed in the INEX 2012 Tweet Contextualization Track consists in obtaining some textual context from the English Wikipedia about the subject of a tweet. The final contextualization of the tweet should take the form of a readable summary of 500 words. An amount of 1133 documents, contextualized tweets with text from Wikipedia from November 2011, were processed in order to obtained summaries. Bibliographic references an empty Wikipedia pages were omitted.

The evaluation of summaries was done by the INEX organizers taking into account informativeness and readability. The former was obtained using Kullback-Leibler divergence with Dirichlet smoothing by comparing n-gram distributions. The latter was accomplished by the participants in the track; they evaluated the summaries taking into account syntax, anaphoric resolution and redundancy. More details of the system of evaluation and the INEX 2012 Tweet Contextualization Track could be found in [1].

For this track we developed a stemmer based on morphological segmentation. The stemmer was coupled with CORTEX, an automatic summarization system, in order to generate the summaries. We tested three sizes of training corpora to determine the best option for statistical stemming for English.

The organization of this paper is as follows: in Section 2 we review some approaches of morphological segmentation; in Section 3 we present word stemming; in Section 4 we describe the affixality measurements; Section 5 presents the stemming strategy; evaluation obtained in INEX track is expose in Section 6 and finally, in Section 7, we briefly present our conclusions and future work.

2 Morphological Segmentation

The first work for unsupervised discovery of morphological units of language is due to Zellig Harris [2]. His method, commonly known as *frequent successor*, consists in counting different letters or symbols before and after a possible morphological boundary. As more different symbols, the probability of a true morphological cut increases. This approach shown, among other things, that uncertainty is a well clue for morphological segmentation.

Now a day, one of the most utilized methods for unsupervised learning of morphology is based on Minimum Description Length (MDL) approach. This has been developed as a computational system called *Linguistica* [3, 4].[1] This method tries to obtain a lexicon of morphs inferred from a corpus. The best lexicon is the one that has the less redundancy, i.e. when the description length of the data is the lowest. Also, this utilizes some combinatorial structures called *signatures* in order to improve segmentation. This method has been employed for stemming work in [5]. In that paper the developed stemmer was utilized for an information retrieval task instead of summarization.

The mission of preprocessing documents for tasks of NLP, such as Question Answering, Information Retrieval or Automatic Text Summarization, in agglutinative languages is more complex. This is due to the fact that agglutinative languages have numerous combinations of morphs rather than a simple *prefix-stem-suffix* combination. A method of unsupervised morphological segmentation for these kinds of languages is called *Morfessor* [6–9]. [2] This approach uses MDL by Maximum a Posteriori framework. Also, it integrates a morphotactic analysis to represent each word by a Hidden Markov Model (HMM). We are not sure if this method has been used for word stemming.

3 Word stemming

The majority of NLP systems preprocesses documents in order to decrease the Vector Space Model representation. This is the case of CORTEX, which will be explained below. A well-known strategy for that purpose is word stemming, i.e.

[1] http://linguistica.uchicago.edu
[2] http://www.cis.hut.fi/projects/morpho/

truncating words by eliminating the inflection. Also, it is possible to remove derivational affixes.

The methods most widely used for word stemming are created by means of hand-made rules, like [10, 11]. These kinds of stemmers have been successfully applied for European languages. However, languages with more complex morphology than English, such as agglutinative ones, need unsupervised morphological strategies in order to deal with language complexity.

In [12] a review of stemming methods is presented. The variety of stemming approaches includes: distance function to measure an orthographical similarity [13], directed graphs [14, 5], and frequency of n-grams of letters [15]. Moreover, there are some works about stemming evaluation in information retrieval tasks, for example [16, 17].

4 Affixality Measurements

The affixality measurements used to morphological segmentation were proposed for Spanish in [18, 19]. These measurements have been also applied to Czech [20], and to the Amerindian Languages Chuj and Tarahumara [21]. This approach lies on the linguistic idea that there is a *force* between segments of a word (morphs) called affixality. If we can quantify this affixality, we can expect some peaks where morphological cuts are possible. In next sections we present the way to calculate these measurements.

4.1 Entropy

As we said above, Harris's approach revealed that uncertainty helps to morphological segmentation. This uncertainty could be seen as the Shannon's concept of information content (entropy) [22]. To calculate the entropy of a possible segmentation, given $a_{i,j}::b_{i,j}$ as a word segmentation, and $B_{i,j}$ as a set of all segments combined with $a_{i,j}$, we can used the formula:

$$H\left(a_{i,j} :: B_{i,j}\right) = -\sum p\left(b_{k,j}\right) \times log_2\left(p\left(b_{k,j}\right)\right) \tag{1}$$

where $k = 1, 2, 3, \ldots |B_{i,j}|$ and each $b_{k,j} \in B_{i,j}$. For our purpose we tested peaks of entropy from right to left in order to discover suffixes.

4.2 Economy Principle

The Economy Principle could be understood as follows: fewer units at one level of language are combined in order to create a great number of other units at the next level. Taking advantage of this principle, we can define a stem as a word segment that belong to a big set of relatively infrequent units, and affixes as word segments that belong to a small set of frequent ones. In [23] a quantification of this economy was suggested, however, we present a reformulation. Given a word

segmentation $a_{i,j}::b_{i,j}$, the economy of a segmentation is calculated depending on type of morph hypothesized:

$$K^p_{i,j} = 1 - \frac{|A_{i,j}| - |A^p_{i,j}|}{|B^s_{i,j}|}; \qquad K^s_{i,j} = 1 - \frac{|B_{i,j}| - |B^s_{i,j}|}{|A^p_{i,j}|} \tag{2}$$

where $A_{i,j}$ is the set of segments which alternate with $b_{i,j}$ ($a_{i,j} \in A_{i,j}$), and $B_{i,j}$ a set of segments which alternate with $a_{i,j}$ ($b_{i,j} \in B_{i,j}$). Also, let $A^p_{i,j}$ be the set of segments which are likely prefixes, and $B^s_{i,j}$ the set of segments which are likely suffixes.

4.3 Numbers of Squares

Joseph Greenberg [24] proposed the concept of square when four expressions of language, let say A, B, C, D, are combined to form AC, BC, AD, and BD. Hence, we set $c_{i,j}$ as a number of squares found in segment j of the word i.

5 Stemming Strategy

The affixality of all possible segmentations within a word is estimated by an average of normalized values of the three explained measurements:

$$AF^n(s_x) = \frac{c_x/\max c_i + k_x/\max k_i + h_x/\max h_i}{3} \tag{3}$$

To calculate this affixality, a training corpus of raw text is required. In this track we use three different sizes of 100k, 200k, and 500k word tokens. With an index of affixality calculated for each possible word segment, it is possible to choose a strategy for morphological segmentation; for example [19] propounded four strategies.

In this experiment we use a peak-valley strategy for segmentation. Given a set of affixality indexes inside a word af^k_i, let $af^k_{i-1} < af^k_i > af^k_{i+1}$ be a peak of affixality from left to right, where k is the length of the word plus one (the ending of the word). The main disadvantage of this approach is that small peaks are taking into account generating oversegmentation.

Regarding stemming, we truncate words at most left peak of affixality. For a language with scare morphology like English, we can imagine that a most right peak of affixality could be sufficient for stemming. However, in order to improve CORTEX summarization, we decide to strongly conflate words by a left-peak strategy. Next section explains CORTEX's approach.

5.1 CORTEX Summarizer

As we mentioned before, CORTEX is an automatic text summarizer system. A wide explanation of this summarizer could be found in [25–29]. Here, we briefly

describe some relevant aspects. First, CORTEX represents input documents in Vector Space Model. To do that, the documents should be preprocessed. Actually, we incorporate our stemmer in this step.

After preprocessing, a frequency matrix γ is generated representing the presence and absence of words (terms) in a sentence:

$$\gamma = \begin{bmatrix} \gamma_1^1 & \gamma_2^1 & \dots & \gamma_i^1 & \dots & \gamma_M^1 \\ \gamma_1^2 & \gamma_2^2 & \dots & \gamma_i^2 & \dots & \gamma_M^2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \gamma_1^P & \gamma_2^P & \dots & \gamma_i^P & \dots & \gamma_M^P \end{bmatrix}, \quad \gamma_i^\mu \in \{0, 1, 2, \dots\} \tag{4}$$

each element γ_i^μ of this matrix represents the number of occurrences of the word$i$ in the sentence μ; $1 \leq i \leq M$ words, $1 \leq \mu \leq P$ sentences.

Then, statistical information is extracted from the matrix by calculating some metrics. More information about these metrics could be found in [30]. A summary of this metrics is offered here; they are based on frequencies, entropy, measures of Hamming and hybrid values.

1. **Frequency measures.**
   (a) Term Frequency: $F^\mu = \sum_{i=1}^{M} \gamma_i^\mu$
   (b) Interactivity of segments: $I^\mu = \sum_{\substack{i=1 \\ \xi_i^\mu \neq 0}}^{M} \sum_{\substack{j=1 \\ j \neq \mu}}^{P} \xi_i^j$
   (c) Sum of probability frequencies: $\Delta^\mu = \sum_{i=1}^{M} p_i \gamma_i^\mu$; p_i = word's i probability
2. **Entropy.** $E^\mu = -\sum_{\substack{i=1 \\ \xi_i^\mu \neq 0}}^{M} p_i \log_2 p_i$
3. **Measures of Hamming.** These metrics use a Hamming matrix H, a square matrix $M \times M$:

$$H_n^m = \sum_{j=1}^{P} \left\{ \begin{matrix} 1 & \text{if } \xi_m^j \neq \xi_n^j \\ 0 & \text{elsewhere} \end{matrix} \right\} \quad \text{for} \quad \begin{matrix} m \in [2, M] \\ n \in [1, m] \end{matrix} \tag{5}$$

   (a) Hamming distances: $\Psi^\mu = \sum_{\substack{m=2 \\ \xi_m^\mu \neq 0}}^{M} \sum_{\substack{n=1 \\ \xi_n^\mu \neq 0}}^{m} H_n^m$
   (b) Hamming weight of segments: $\phi^\mu = \sum_{i=1}^{M} \xi_i^\mu$
   (c) Sum of Hamming weight of words per segment: $\Theta^\mu = \sum_{\substack{i=1 \\ \xi_i^\mu \neq 0}}^{M} \psi_i$; every word. $\psi_i = \sum_{\mu=1}^{P} \xi_i^\mu$
   (d) Hamming heavy weight: $\Pi^\mu = \phi^\mu \Theta^\mu$
   (e) Sum of Hamming weights of words by frequency: $\Omega^\mu = \sum_{i=1}^{M} \psi_i \gamma_i^\mu$
4. **Titles.** $\theta^\mu = \cos\left(\frac{\sum_{i=1}^{M} \gamma_i^\mu \text{Title}}{\|\gamma^\mu\| \|\text{Title}\|} \right)$

Finally, a decision algorithm combines those metrics to score sentences. Two averages are calculated, $\lambda_\mu > 0.5$, and $\lambda_\mu < 0.5$ ($\lambda_\mu = 0.5$ is ignored):

$$\sum^{\mu} \alpha = \sum_{\substack{\nu=1 \\ \|\lambda_\mu^\nu\| > 0.5}}^{\Gamma} \left(\|\lambda_\mu^\nu\| - 0.5 \right); \sum^{\mu} \beta = \sum_{\substack{\nu=1 \\ \|\lambda_\mu^\nu\| < 0.5}}^{\Gamma} \left(0.5 - \|\lambda_\mu^\nu\| \right) \tag{6}$$

The next expression is used to calculate the score of each sentence:

$$\text{If } \left(\sum^{\mu} \alpha > \sum^{\mu} \beta \right)$$
$$\text{then } \Lambda^{\mu} = 0.5 + \frac{\sum^{\mu} \alpha}{\Gamma} \text{ else } \Lambda^{\mu} = 0.5 - \frac{\sum^{\mu} \beta}{\Gamma}$$

CORTEX sorts final sentences by using $\Lambda^{\mu}; \mu = 1, \cdots, P$. Additionally, CORTEX let us delimit a compression rate, which was fixed at 500 words.

6 Experiments and Results

6.1 Design of Experiments

We made use of three sizes of training corpora, 100K, 200K, and 500K word tokens, to test our stemmer. With these sizes we performed the three runs for INEX track. The assigned numbers of runs were 153 (100K), 154 (200K), and 155 (500K). The corpus for evaluation was the 1133 contextualized tweets with text from Wikipedia from November 2011. About training corpora, we selected 24 documents from the same contextualized tweets.

6.2 Results

For informativeness, CORTEX, coupled with our stemmer, obtained rank 12, 14, and 15. Average scores of informativeness are shown in Table 1. The best run in this evaluation was run 154 (200K).

Table 1. Average scores of informativeness

Rank	Run	Unigrams	Bigrams	Skip
12	**154**	**0.8233**	**0.9254**	**0.9251**
14	155	0.8253	0.9280	0.9274
15	153	0.8266	0.9291	0.9290

Those scores were computed by organizers using a Perl script (inexqa-eval.pl); for details about this script check [1].

On the other hand, the best results for readability evaluation were obtained by run 155 (500K), see Table 2. Comparing our results with other runs, run 155 (500K) obtained rank 4 in relevance, rank 6 in syntax, and rank 9 in structure. The worst run in our experiment was the run 153 (100K) in both evaluations.

Table 2. Scores of readability

Run	Relevance	Syntax	Structure
155	**0.6968**	**0.6161**	**0.5315**
154	0.5352	0.5305	0.4748
153	0.4984	0.4576	0.3784

7 Conclusions and Future Work

In this paper we reported an experiment using a stemmer based on morphological segmentation. We used affixality measurements in order to segment words. This stemmer was coupled with CORTEX, an automatic summarization system.

We suggested the next stemming strategy: given some peaks of affixality of a word, we truncated at most left peak. Also, we tested three training corpus sizes to obtain statistical information for the affixality indexes: 100K, 200K, and 500K word tokens. Our two goals were to know if our stemming strategy can produce readable summaries, and if different sizes of training corpora can improve the CORTEX performance.

According to results of evaluation, our stemming strategy produces not only readable summaries but also competitive ones. That is, from an average of relevance, syntax, and structure (0.6148), run 155 obtained a rank 7 among 27 runs. What is more, concerning informativeness, run 154 obtained rank 12 among 33 participants.

Regarding corpus sizes, it is not clear what size is the best for English, between 200K and 500K word tokens. However, it is clear that increasing corpus size is a good strategy because 100K obtained the worst results. Additionally, a greater training corpus gives better position in the ranking, for example, from an average of relevance, syntax, and structure, run 155 (500K) obtained rank 7 and run 153 (100K) obtained rank 15.

In future experiments we will test different strategies for morphological segmentation and stemming. Additionally, we can test different stemming approaches, such as Porter's stemmer.

References

1. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the INEX 2011 Question Answering Track (QA@INEX). In: INEX 2011 Workshop Pree-Proceedings, IR Publications, Hofgut Imsbach, Saarbrücken, Germany (2011) 145–153
2. Harris, Z.S.: From Phoneme to Morpheme. Language **31** (1955) 190–222
3. Goldsmith, J.: Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics **27** (2001) 153–198
4. Goldsmith, J.: An Algorithm for the Unsupervised Learning of Morphology. Natural Language Engineering **12** (2006) 353–371

5. Paik, J.H., Mitra, M., Parui, S.K., Jarvelin, K.: GRAS: An effective and efficient stemming algorithm for information retrieval. ACM Trans. Inf. Syst. **29** (2011)
6. Creutz, M., Lagus, K.: Unsupervised Discovery of Morphemes. In: Proc. of the Workshop on Morphological and Phonological Learning of ACL-02, Philadelphia, SIGPHON-ACL (2002) 21–30
7. Creutz, M.: Unsupervised segmentation of words using prior distributions of morph length and frequency. In Hinrichs, E., Roth, D., eds.: 41st Annual Meeting of the ACL, Sapporo, Japan. (2003) 280–287
8. Creutz, M., Lagus, K.: Induction of a Simple Morphology for Highly-Inflecting Languages. In: Proc. of 7th Meeting of the ACL Special Interest Group in Computational Phonology SIGPHON-ACL. (2004) 43–51
9. Creutz, M., Lagus, K.: Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In: Int. and Interdisciplinary Conf. on Adaptive Knowledge Representation and Reasoning (AKRR05). (2005) 106–113
10. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics **11** (1968) 23–31
11. Porter, M.F.: An algorithm for Suffix Stripping. Program **14** (1980) 130–137
12. Lennon, M., Pierce, D., Tarry, B., Willet, P.: An evaluation of some conflation algorithms for information retrieval. J. of Information Science **3** (1981) 177–183
13. Majumder, P., Mitra, M., Pal, D.: Bulgarian, Hungarian and Czech stemming using YASS. In: Proceedings of Advances in Multilingual and Multimodal Information Retrieval, Springer-Verlag, Berlin (2008) 49–56
14. Bacchin, M., Ferro, N., Melucci, M.: A probabilistic model for stemmer generation. Mechanical Translation and Computational Linguistics **41** (2005) 121–137
15. McNamee, P., Mayfield, J.: Character n-gram tokenization for European language text retrieval. Information Retrieval **7** (2004) 73–97
16. Krovetz, R.: Viewing Morphology as an Inference Process. In: Proccedings of the 16th ACM/SICIR Conference. (1993) 191–202
17. Hull, D.A.: Stemming algorithms - A case study for detailed evaluation. Journal of the American Society for Information Science **47** (1996) 70–84
18. Medina-Urrea, A.: Investigación cuantitativa de afijos y clíticos del español de México. Glutinometría en el Corpus del Español Mexicano Contemporáneo. PhD thesis, El Colegio de México, México (2003)
19. Medina-Urrea, A.: Automatic Discovery of Affixes by means of Corpus: A Catalog of Spanish Affixes. Journal of Quantitative Linguistics **7** (2000) 97–114
20. Medina-Urrea, A., Hlaváčová, J.: Automatic Recognition of Czech Derivational Prefixes. In: Proceedings of CICLing 2005. Volume 3406., LNCS, Springer, Berlin/Heidelberg/New York (2005) 189–197
21. Medina-Urrea, A.: Affix Discovery based on Entropy and Economy Measurements. Texas Linguistics Society **10** (2008) 99–112
22. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
23. de Kock, J., Bossaert, W.: Introducción a la lingüística automática en las lenguas románicas. Gredos, Madrid (1974)
24. Greenberg, J.H.: Essays in Linguistics. The Univ. of Chicago Press, Chicago (1957)
25. Torres-Moreno, J.M.: Résume automatique de documents. Lavoisier, Paris (2011)
26. Torres-Moreno, J.M., Saggion, H., da Cunha, I., SanJuan, E., Velázquez-Morales, P.: Summary Evaluation with and without References. Polibits **42** (2010) 13–19
27. Saggion, H., Torres-Moreno, J.M., da Cunha, I., SanJuan, E.: Multilingual summarization evaluation without human models. In: 23rd Int. Conf. on Computational Linguistics. COLING '10, Beijing, China, ACL (2010) 1059–1067

28. Torres-Moreno, J.M., Velazquez-Moralez, P., Meunier, J.: CORTEX, un algorithme pour la condensation automatique de textes. In: ARCo. Volume 2. (2005) 365
29. Torres-Moreno, J.M., Velazquez-Morales, P., Meunier, J.: Condensés de textes par des méthodes numériques. JADT **2** (2002) 723–734
30. Torres-Moreno, J.M., St-Onge, P.L., Gagnon, M., El-Bèze, M., Bellot, P.: Automatic Summarization System coupled with a Question-Answering System (QAAS). CoRR **abs/0905.2990** (2009)