Extrinsic Evaluation on Automatic Summarization Tasks: Testing Affixality Measurements for Statistical Word Stemming

Carlos-Francisco Méndez-Cruz[1,4], Juan-Manuel Torres-Moreno[1,2], Alfonso Medina-Urrea[3], and Gerardo Sierra[4]

[1] LIA-Université d'Avignon et des Pays de Vaucluse, France
cmendezc@ii.unam.mx, juan-manuel.torres@univ-avignon.fr
[2] École Polytechnique de Montréal, Canada
[3] El Colegio de México A.C., México
amedinau@colmex.mx
[4] GIL-Instituto de Ingeniería UNAM, México
gsierram@ii.unam.mx

Abstract. This paper presents some experiments of evaluation of a statistical stemming algorithm based on morphological segmentation. The method estimates affixality of word fragments. It combines three indexes associated to possible cuts. This unsupervised and language-independent method has been easily adapted to generate an effective morphological stemmer. This stemmer has been coupled with CORTEX, an automatic summarization system, in order to generate summaries in English, Spanish and French. Summaries have been evaluated using ROUGE. The results of this extrinsic evaluation show that our stemming algorithm outperforms several classical systems.

Keywords: Automatic summarization, Affixality Measurements, Morphological Segmentation, Statistical Stemming, CORTEX.

1 Introduction

Discovering linguistic units in Natural Languages has been a long-standing human task. Now, automatic approaches are used to conduct this work. Despite rule-based approaches for word processing are widely used, there is a renewed interest in morphological methods. Corpora of morphologically complex languages, such as agglutinative ones, can be processed in order to discover morphemes. Simple strategies are not suitable for this kind of languages. Also, corpora for languages, which have not been computationally studied, appear every day. Then, this paper presents a method for unsupervised learning of morphology.

A morphologist tries to collect morphological units from languages, among other tasks. Roughly, two types of units are considered in this paper: stems and affixes. On the one hand, a stem is the part of a word stripped of affixes, it carries lexical content. On the other hand, affixes carry either grammatical information or information to produce another word. Finally, by its position, affixes can be

I. Batyrshin and M. González Mendoza (Eds.): MICAI 2012, Part II, LNAI 7630, pp. 46–57, 2013.

classified in prefixes, before the stem; and suffixes, following it. In morphologically complex languages, words can be recursively formed by stems and affixes. Thus, this paper deals with discovering stems and suffixes, i.e. morphological segmentation and its evaluation. We propose an extrinsic evaluation by means of a Natural Language Processing (NLP) task. We developed a stemmer to test some procedures to eliminate suffixes from a word. The stemmer is coupled with an automatic text summarizer. Lastly, summaries are evaluated with ROUGE, a semi-automatic evaluation system.

2 State of the Art

2.1 Morphological Segmentation

Morphological segmentation tries to discover morphological units (morphs) from a sequence of symbols of language. This problem is closely dependent on language. Nonetheless, computational approaches have made traditionally simplistic assumptions such as a word has a simple structure as *prefix-stem-suffix* [1].

The first work for unsupervised discovery of morphemes is due to Zellig Harris [2]. His approach consisted in counting distinct symbols preceding/following a possible morphological boundary. Thus, high frequency corresponds to true morphological boundaries. After Harris's work, many approaches have been proposed. A wide review of methods can be seen in [3,4,5]. A large majority of them sees morphology as a lexicon of words. Their aim is to reduce the redundancy in the lexicon. A method that follows this idea was proposed in [6,7]. This employs Minimum Description Length (MDL) analysis as a strategy for obtaining the lowest redundancy in the lexicon. Hence, the best morphology is obtained when the description length of the data is the lowest. To control the quality of segmentations, this method uses combinatorial structures named *signatures*. Also, signatures are involved in calculating the description length. This approach has been implemented in a computational program named Linguistica.[1]

Another method that searches also an optimal morphological model has been presented in [8,1,9,10]. The set of methods developed in these papers has been called *Morfessor*. It has been formulated for agglutinative languages such as Finnish. Regarding [10], it incorporates a morphotactic analysis for each word. Thus, given some initial segmentation, four possible categories are assigned: prefix, suffix, stem, and no-morph. The method uses Maximum a Posteriori (MAP) framework, which is essentially equivalent to MDL. In order to assign those categories, each word is represented by a Hidden Markov Model (HMM).

An effort for obtaining a minimal model of inflectional morphology for Spanish is exposed in [11,12] It uses genetic algorithms for finding the best lexicon. The fitness function is based on MDL principles.

---

[1] `http://linguistica.uchicago.edu`

2.2 Stemming Works

In general, stemming algorithms aim to remove inflectional and derivational suffixes from words. In many tasks of NLP, such as information retrieval, question-answering, or automatic summarization, stemming is an important part of text preprocessing. Often, a document is represented as a Vector Space Model. Then, in order to improve performance, conflating similar words is preferred.

Thus, language dependent approaches based on handmade rules are widely used, for instance [13]. In fact, the Porter's stemmer [14] is utilized in many NLP systems for European languages. In this approach, a set of removing rules is proposed, where suffixes are substituted by other ones, even the null suffix. A short example from [14] (Step 3) is listed in (1).

(1) a. ICITI → IC (electriciti → electric)
  b. ICAL → IC (electrical → electric)
  c. FUL → null (hopeful → hope)

However, information requirements in more languages have emerged. Also, agglutinative languages need strategies for stemming different from simply suffix removal. These facts hold research attention in unsupervised approaches, instead of adapting Porter's rules. Furthermore, some works have focused on the disadvantages of rule-based approaches, like [15].

A review of stemming methods can be found in [16]. A language independent approach [17] presents a clustering based unsupervised technique named YASS. A distance function is used in order to measure orthographical similarity, and to assign a low distance value to related words. Other methods which use graphs are [18,19]. The former splits words from a text at all possible positions. Then, segmentations are represented as a directed graph. An algorithm is used to estimate prefix and suffix scores to maximize the probability of a prefix-suffix pair. On the other hand, a method that uses frequency of n-grams of letters as strategy of stemming is presented in [20].

A word regularization process close to stemming is lemmatization. Given a group of words grammatically related, this process selects a representative of them (lemma). However, [21] has stressed some problems in NLP tasks. Also, it presents an unsupervised algorithm based on word clustering by means of a similarity matrix that searches a lemma for a group of words semantically related. Finally, there has been a strong interest in stemming evaluation. For instance, [15] made an evaluation of some stemming algorithms for information retrieval. This work showed that stemming improves performance significantly in short documents. We found similar conclusions in [22], where five stemming algorithms were evaluated using the SMART text retrieval system.

3 Affixality Measurements

A brief description of affixality measurements is presented in this section. In [5,23] these measurements are exposed with more detail for Spanish. Also, its application in unrelated languages could be found in [24] for Czech, and [25] for the Amerindian Languages Chuj and Tarahumara. The idea behind this approach is to quantify the affixality of a word segment. In other words, it tries to estimate the combinatorial *force* of linguistics units. One expects that higher values of that affixality correspond to morphological segmentations. It is clear that, regarding other methods, this one is not guided for searching an optimal morphological model. Actually, three statistical measurements are used for computing the affixality. They are presented below.

3.1 Entropy

Harris's idea revealed that uncertainty is a good indicator of morphological cuts. This idea is closely related to Shannon's concept of information content (entropy) [26]. Therefore, given $a_{i,j}{::}b_{i,j}$ as a word segmentation,[2] and $B_{i,j}$ a set of all segments combined with $a_{i,j}$ the entropy is obtained as follows:

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

where $k = 1, 2, 3, \ldots |B_{i,j}|$ and each $b_{k,j} \in B_{i,j}$. As it was tested in [23], peaks of entropy from right to left are good indicators of a suffix segmentation.

3.2 Economy Principle

Morphological phenomena work as economical systems. Fewer units are combined at one level in order to create a great number of another units at the next level (Economy Principle). Thus, we can expect that word stems belong to a big set of relatively infrequent units, and affixes to a small set of frequent ones.

In [27] a quantification of this economy was proposed. Here, a reformulation is presented. Given a word segmentation $a_{i,j}{::}b_{i,j}$, let $A_{i,j}$ be 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 $a_{i,j}$ ($b_{i,j} \in B_{i,j}$). Now, 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. Thus, the economy of a segmentation is formulated in two ways, depending on type of morph hypothesized:

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

3.3 Numbers of Squares

A square is found in language when four expressions, let say A, B, C, D, are combined to form AC, BC, AD, and BD. This concept was proposed by Joseph Greenberg in [28]. An example in Spanish would be *abarc::aba*, *camin::aba*, *abarc::aron*, *camin::aron*. Hence, let $c_{i,j}$ be the number of squares found in segment j of the word i.

[2] We use :: as a segmentation mark.

3.4 Measurements Combined

Then, affixality is estimated by a combination of the three measurements previously explained. Consequently, an average of normalized values is calculated as a combination strategy:

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

An important fact is that no explicit distinction between inflection and derivation is involved in the procedure of affixality calculation. In addition, affixality could be obtained either from left to right or conversely. Left to right affixality permits us to discover prefixes, while right to left affixality fits in better with suffixes. For calculating these measurements only raw text is necessary as a training corpus. In our experiments we will evaluate three differents sizes (see section 5.1). An example of the affixality index for a Spanish word is shown in Table 1. The affixality index has been used in previous work for gathering affix catalogs [5,23]. The idea is to split a word in two segments, taking the highest value as the cut. For instance, in Table 1 the cut occurs between the stem UTILIZADO∼ and the suffix ∼S (UTILIZADO::S). Our first interest is proposing a segmentation strategy which leads us to discover all morphs (in this paper suffixes).

Table 1. Affixality for Spanish word: `UTILIZADOS` (masculine and plural form of past participle of verb TO UTILIZE)

U	**T**	**I**	**L**	**I**	**Z**	**A**	**D**	**O**	**S**
« 0.079	0.0698	0.1291	0.1566	0.3066	0.7137	0.1556	0.8097	0.8231	

« Right to left affixality

4 Methodology

In this section, we describe a simple approach, based on the affixality index to morphological segmentation. Then, we propose an extrinsic evaluation strategy by state-of-art automatic summarizer.

4.1 Morphological Segmentation by Affixality Measurements

Once the affixality index has been obtained, we can choose different ways to discover morphological segments. Regarding that, [23] propounded four strategies: (1) *max(affixality)*, (2) *affixality > a threshold*, (3) *affixality > 0*, and (4) *max(affixality) recursively*. Here, we use a simple peak-valley strategy for segmentation. Given a set of affixality indexes inside a word af_i^k, let $af_{i-1}^k < af_i^k > af_{i+1}^k$ be a peak of affixality from left to right, where k is the length of the word plus one: the ending of the word.

Furthermore, we used two heuristic strategies. First, we start searching for a peak of affixality at $i = 3$. Second, the ending of a word has zero affixality, i.e. $af_k = 0$. The first assumption let us avoid stem over-segmentation, whereas the second assumption allows one-letter suffixes at the end of a word. For instance, from Table 1 we obtained the segmentation UTILIZ::ADO::S. There are some disadvantages to this approach. For example, slight peaks could be taking into account as possible morphological cuts. Once the new morphological segmentation strategy is proposed, an evaluation will be required. There are two criteria for evaluation: intrinsic and extrinsic [29]. Intrinsic evaluation for morphological segmentation requires comparison to a morphological gold standard. Despite it is possible to find some handmade segmentation corpus, we suggest for this paper an extrinsic evaluation.

A natural purpose for morphological segmentation is stemming. In [30] highest affixality values were suggested for word truncation. The proposed evaluation showed that this approach outperforms the Porter's stemmer. In that paper, the affixality was calculated only with economy and entropy measurements. In this paper, we additionally use squares. Alternatively, our goal consists in testing two strategies that generate two possible cuts. First, truncating at the rightmost peak of the word (UTILIZ::ADO::S > UTILIZADO∼) would result in inflectional stemming. Second, truncating at leftmost peak (UTILIZ::ADO::S > UTILIZ∼) would result in derivational-inflectional stemming.

4.2 Automatic Summarization Like Extrinsic Evaluation Task

An automatic summarization task was selected for this extrinsic evaluation. Particularly, the Cortex system was chosen for this purpose. Its modular architecture allowed us to adapt easily the new stemmer. Preprocessing texts is the first step in Cortex. Texts are filtered, and words are lemmatized by means of a dictionary. The number of words in the dictionary is for English 97K, for Spanish 624K and for French 194K. Instead of this lemmatization stage, the Porter's stemmer could be used.[3] Even both processes could be sequentially applied. In this step, our stemmer has been coupled with Cortex.

4.3 Cortex Summarizer

An automatic text summarizer extracts the most relevant sentences from a text [31]. The Cortex system generates summaries from texts represented in Vector Space Model. A decision algorithm combines several metrics in order to score sentences. After filtering, a frequency matrix γ is constructed: 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. Matrix ξ represents the presence/absence of terms in a sentence μ.

[3] Specifically, a Perl implementations from Lingua-Stem-Snowball, `http://snowball.tartarus.org/`

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

4.4 The Metrics

Important mathematical and statistical information can be extracted from the matrices γ and ξ. CORTEX uses Γ metrics calculated on frequencies, entropy, Hamming and hybrid values. See [32] for more information of CORTEX's metrics.

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 $M \times M$:

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

   (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.** $\theta^\mu = \cos\left(\frac{\sum_{i=1}^{M} \gamma_i^\mu \text{Title}}{\|\gamma^\mu\| \|\text{Title}\|} \right)$

4.5 Decision Algorithm (DA)

The Decision Algorithm combines all normalized sentence scores. Two averages are calculated: $\lambda_\mu > 0.5$, and $\lambda_\mu < 0.5$ ($\lambda_\mu = 0.5$ is ignored):

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

ν is the index of the metrics, $\sum_{\nu}^{\Gamma}$ is the sum of the absolute differences between $\|\lambda\|$ and 0.5, $\sum^{\mu}\alpha$ are the "positive" normalized metrics, $\sum^{\mu}\beta$ the "negative" normalized metrics and Γ the number of metrics used. The score of each sentence is calculated as follows:

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

All sentences are sorted using $\Lambda^{\mu}; \mu = 1, \cdots, P$. The compression rate τ determines the final size of the summary.

5 Experiments and Results

5.1 Design of Experiments

In order to evaluate our stemmer, we performed some experiments involving (i) three different stemming strategies, (ii) corpora in different languages: English, Spanish and French, and (iii) different sizes of training corpora: 100K, 200K, and 500K word tokens. Our stemmer is compared to CORTEX's lemmatizer (lemm), the Porter's stemmer used by CORTEX (stem), and both methods sequentially applied (lems). In addition, we included no stemming at all (raw) and fixed truncation at 6 characters (fixed). Respecting to our stemmer, we tested three possible cuts of a word based on morphological segmentation, all of them with the three sizes of training corpora: highest affixality value (vM100, vM200, vM500), first peak of affixality at right (R100, R200, R500), and first peak of affixality at left (L100, L200, L500).

Summaries were evaluated using ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [33]. This semi-automatic evaluation system calculates a score of similarity between a candidate summary and several human summaries. We evaluated them using bigrams (ROUGE-2) and skip bigrams (ROUGE-SU4).

5.2 Corpora

Three sets of documents were selected for the evaluation in English, Spanish and French. For English, 50 clusters from DUC 2004[4] were used, specifically the *Task 2 - Short multi-document summaries focused by TDT events*. The clusters contained on average 10 documents from the AP newswire and the New York Times. Four human summaries per cluster were available for summaries evaluation. For Spanish, 8 biomedical articles were collected from the specialized journal *Medicina Clínica*.[5] In this case, we evaluated automatic summaries against author's

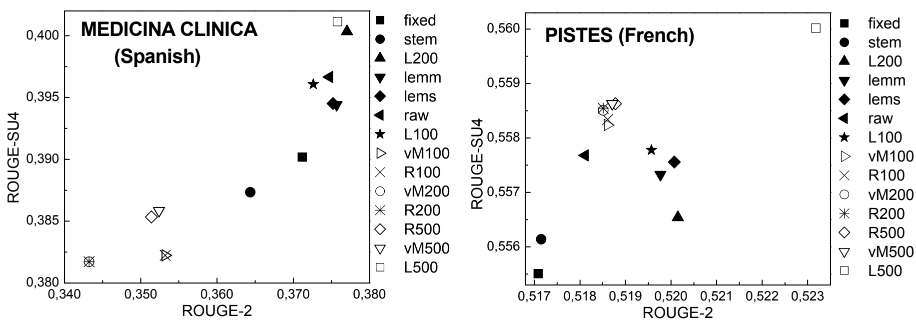
[4] http://duc.nist.gov/duc2004
[5] http://www.elsevier.es/revistas/ctl_servlet?_f=7032&revistaid=2

abstracts. Regarding French evaluation, we utilized *Canadien French Sociological Articles* corpus [34] from the specialized e-journal *Perspectives interdisciplinaires sur le travail et la santé* (PISTES).[6] 50 sociological articles constituted this corpus. One human abstract per text was used for evaluation.

Regarding training corpora, we used different texts than those in the evaluation task. We constituted three sets of documents of 100K, 200K and 500K word tokens per each language. For English, we selected 24 documents from INEX 2012 *Tweet Contextualization Track*.[7] Each document was formed by passages from Wikipedia that contextualized a tweet query. For Spanish, we used the *Corpus del Español Mexicano Contemporáneo, CEMC* [35], a well-balanced text collection from different sources. Finally, for the French corpus we gathered texts from several sources. First, two corpora from *DÉfi Fouille de Textes* (DEFT'07) were included. Second, we used *aVoiraLire* (3 460 critiques of books, films, comics and shows) and *relectures* (1 484 reviews of scientific articles). Also, the book *Pensée, essais, maximes et correspondance de J. Joubert* was included. At last, 2 472 phrases of texts of Jules Verne were considered.

5.3 Results

The ROUGE metrics obtained on the three corpora are plotted in Figures 1 and 2. For Spanish (Figure 1(a)) and French (Figure 1(b)), we can observe that the method L500 obtained the best score.



(a) Spanish corpus (b) French corpus

Fig. 1. Content evaluation for Spanish and French corpus

On the other hand, first peak of affixality at right and the highest value of affixality methods obtained lower scores. It means that CORTEX performs well for Spanish and French with a derivational-inflectional stemmer. By comparison, this method (L500) was the worst on English corpus (see Figure 2). In fact, the

[6] http://www.pistes.uqam.ca/index.html
[7] https://inex.mmci.uni-saarland.de/tracks/qa/

truncating methods at the end of the word appeared more successful, for instance vM100, R100. This suggests a correlation between the segmentation approach and the morphological complexity.

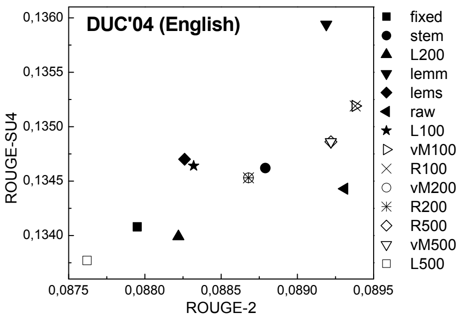


Fig. 2. Content evaluation for English corpus

Surprisingly, the raw method performed as good as CORTEX's strategies (lemm, lems), and overcame stem and some affixality approaches for Spanish. Besides, in this language also the fixed method was better than the Porter's stemmer. Similar situation was found in French (see Figure 1(b)), where the raw method defeated the stem method. Consequently, CORTEX works better while word normalization increases. This could explain the first position of lemm (a dictionary method) for English.

Regarding corpus sizes, position of L500 in Spanish and French evaluation proposes that an improvement is obtained increasing the size of the training corpus. Nevertheless, for English it is not true. We can explain this because of the relatively simple inflectional morphology of this language. Lastly, we can observe for all languages coincidences between the highest affixality method and first peak method of affixality at right. This situation is due to the fact that the most affixality segments are inflection suffixes, in fact, the leftmost suffixes.

6 Conclusions and Future Work

We have presented a study of statistical stemming through morphological segmentation. This unsupervised and language-independent approach uses affixality measurements for unsupervised learning of morphology. Our results confirm a correlation between the stemming strategy and the morphological complexity of a language. It means that the rule-based stemming loses effectiveness as the morphological complexity increases. In this manner, our statistical stemmer outperforms, for Spanish and French, Porter's and based-dictionary stemming strategies. In consequence, morphological stemmers should be taken into account.

Regarding morphological segmentation, there is always room for some improvement. Different strategies for segmenting should be tested. Even though, results of extrinsic evaluation by means of automatic summarization seem to be promising, an intrinsic evaluation against a gold standard should be done. In the future we will evaluate our method by using FRESA [34,36], it will allow us to test in different corpora avoiding having human summaries.

Acknowledgments. We thank El Colegio de México A.C. for allowing us to use the CEMC corpus. This work was partially supported by a CONACyT grant.

References

1. Creutz, M., Lagus, K.: Unsupervised Discovery of Morphemes. In: Proc. of the Workshop on Morphological and Phonological Learning of ACL 2002, Philadelphia, SIGPHON-ACL, pp. 21–30 (2002)
2. Harris, Z.S.: From Phoneme to Morpheme. Language 31, 190–222 (1955)
3. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Trans. Speech Lang. Process 4 (2007)
4. Goldsmith, J.A.: Segmentation and Morphology. In: The Handbook of Computational Linguistics and Natural Language Processing, pp. 364–393. Wiley-Blackwell, Oxford (2010)
5. 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)
6. Goldsmith, J.: Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics 27, 153–198 (2001)
7. Goldsmith, J.: An Algorithm for the Unsupervised Learning of Morphology. Natural Language Engineering 12, 353–371 (2006)
8. 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, pp. 280–287 (2003)
9. 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, pp. 43–51 (2004)
10. 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 (AKRR 2005), pp. 106–113 (2005)
11. Gelbukh, A., Alexandrov, M., Han, S.-Y.: Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model. In: Sanfeliu, A., Martínez Trinidad, J.F., Carrasco Ochoa, J.A. (eds.) CIARP 2004. LNCS, vol. 3287, pp. 432–438. Springer, Heidelberg (2004)
12. Reyes, D.: Sistema de segmentación automática de palabras para el español. Master's thesis, CIC-IPN (2008)
13. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics 11, 23–31 (1968)
14. Porter, M.F.: An algorithm for Suffix Stripping. Program 14, 130–137 (1980)
15. Krovetz, R.: Viewing Morphology as an Inference Process. In: Proccedings of the 16th ACM/SICIR Conference, pp. 191–202 (1993)

16. Lennon, M., Pierce, D., Tarry, B., Willet, P.: An evaluation of some conflation algorithms for information retrieval. J. of Information Science 3, 177–183 (1981)
17. Majumder, P., Mitra, M., Pal, D.: Bulgarian, Hungarian and Czech Stemming Using YASS. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 49–56. Springer, Heidelberg (2008)
18. Bacchin, M., Ferro, N., Melucci, M.: A probabilistic model for stemmer generation. Mechanical Translation and Computational Linguistics 41, 121–137 (2005)
19. 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)
20. McNamee, P., Mayfield, J.: Character n-gram tokenization for European language text retrieval. Information Retrieval 7, 73–97 (2004)
21. Torres-Moreno, J.M.: Reagrupamiento en familias y lexematización automática independientes del idioma. Inteligencia Artificial 47, 38–53 (2010)
22. Hull, D.A.: Stemming algorithms - A case study for detailed evaluation. Journal of the American Society for Information Science 47, 70–84 (1996)
23. Medina-Urrea, A.: Automatic Discovery of Affixes by means of Corpus: A Catalog of Spanish Affixes. Journal of Quantitative Linguistics 7, 97–114 (2000)
24. Medina-Urrea, A., Hlaváčová, J.: Automatic Recognition of Czech Derivational Prefixes. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 189–197. Springer, Heidelberg (2005)
25. Medina-Urrea, A.: Affix Discovery based on Entropy and Economy Measurements. Texas Linguistics Society 10, 99–112 (2008)
26. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
27. de Kock, J., Bossaert, W.: Introducción a la lingüística automática en las lenguas románicas. Gredos, Madrid (1974)
28. Greenberg, J.H.: Essays in Linguistics. The Univ. of Chicago Press, Chicago (1957)
29. Spärck-Jones, K., Galliers, J.: Evaluating Natural Language Processing Systems: An Analysis and Review. Springer, New York (1996)
30. Medina-Urrea, A.: Towards the Automatic Lemmatization of 16th Century Mexican Spanish: A Stemming Scheme for the CHEM. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 101–104. Springer, Heidelberg (2006)
31. Torres-Moreno, J.M.: Résume automatique de documents, Lavoisier, Paris (2011)
32. 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)
33. Lin, C.Y.: Rouge: A Package for Automatic Evaluation of Summaries. In: Workshop on Text Summarization Branches Out (2004)
34. 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 2010, pp. 1059–1067. ACL, Beijing (2010)
35. Lara, L., Ham Chande, R., García Hidalgo, M.: Investigaciones lingüísticas en lexicografía. El Colegio de México, A.C., México (1979)
36. Torres-Moreno, J.M., Saggion, H., da Cunha, I., SanJuan, E., Velázquez-Morales, P.: Summary Evaluation with and without References. Polibits 42, 13–19 (2010)