Natural Language Engineering **19** (4): 537–555. © Cambridge University Press 2012
doi:10.1017/S1351324912000071

On morphological relatedness

AHMED KHORSI

College of Computer and Information Science, Al-Imam Mohammad Ibn Saud Islamic University, Riyadh, Kingdom of Saudi Arabia
email: amakhorsi@imamu.edu.sa,ahmed_khorsi@yahoo.fr

(*Received 28 March 2011; revised 9 January 2012; accepted 9 January 2012; first published online 10 February 2012*)

Abstract

In this paper, we discuss the results of a new unsupervised and computationally lightweight scoring of how two words are morphologically related to each other. This measure is meant to be an alternative to stemming, radicals (root) extraction, and morphological analysis in a wide range of applications; especially information extraction related ones. Compared to light stemming, which seems to be the most convenient approach for systems with efficiency concerns, our measure does not neglect unconditionally a prefix or a suffix as the light stemming does. Instead, our measure takes into account all letters of the word but with different weights. This prevents the missing of a significant letter. Compared to heavy stemming, morphological analysis, or radicals extraction, which rely on dictionaries and compatibility databases, our measure does not rely on any language-specific morphology knowledge. This makes our approach unsupervised and theoretically language independent and computationally much lighter. Our tests targeted Arabic: a Semitic language recognized to have a complex morphology due to its highly inflectional lexicon.

1 Introduction

One of the challenges in textual information extraction is word variants normalization. This task becomes harder for inflectional languages such as Arabic.

In the concatenative languages (Sproat 1992; Menn 1995) like English, words are generally in the form *prefix.stem.suffix* and the problem is then to find frontiers (i.e., the starting letter of *stem* and *suffix*). In more complex morphologies, for instance Arabic (Soudi and van den Bosch 2007; Jurafsky and Martin 2009), the stem could not be considered as the normal carrier of the word's and variants' broad meaning. This is due to the internal changes that affect the word with no substantial effect on the meaning. For instance, stemming the plural form 'كتب' ([kutub]: books)(IPA 1999) will not allow the match of either the singular form 'كتاب' ([kɪtɑːb]: book) or the smallness form 'كتيب' ([kutajɪb]: small book). Both have an extra internal letter: respectively 'ـا' ([ɑː]: 1^{st} Arabic letter) and 'ـيـ' ([j]: 28^{th} Arabic letter) at the third position of each.

Plural: ك ـتـ ـب
Singular: ك ـتـ ـا ـب
Smallness: ك ـتـ ـيـ ـب

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

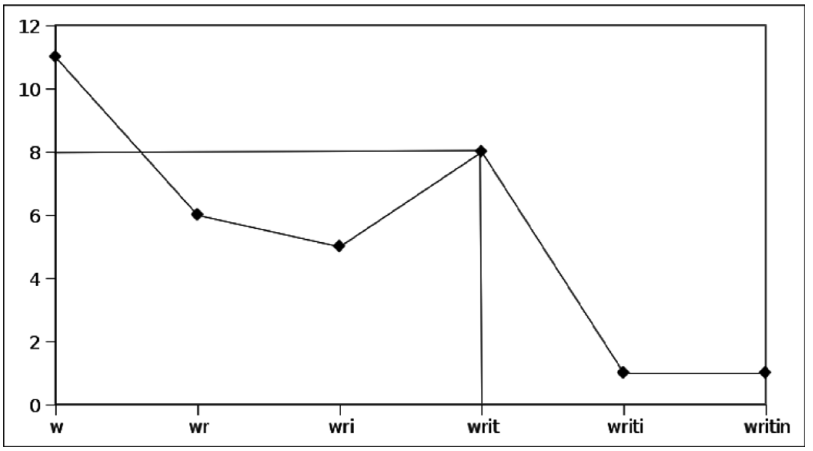


Fig. 1. Successors counts variation for the word '*writing*'.

Arabic is recognized to be highly inflectional and classical stemming techniques are not so effective and perform much worse than they do in concatenative languages.

Being language independent, unsupervised approaches seem to be the oldest candidates to handle non-European languages such as Arabic.

The oldest published work on unsupervised word segmentation is due to Zelig Harris who suggested in the 1950 s a process to break a phonetic text into morphemes (Harris 1955) using only successors numbering. The principle is to count at each position of the word the number of successors and/or predecessors observed in a corpus; positions where the number of predecessors/successors exceeds a threshold k are marked as frontiers.

Example 1.1

Figure 1 shows the curve of successors counts for the word '**writing**.' The curve rises for the substring '*writ*' which is actually a stem.

Figure 2 shows the curve of successors counts of the Arabic word 'والكتاب' ([welkɪtɑ:b]: and the book) observed in a spelling dictionary. Successors number of the prefix 'وال' ([wel]: and the) is a clear peak.

Hafer and Weiss (Hafer and Weiss 1974) adapted Harris's technique to word segmentation and tested the use of fifteen different setups of four measures:

(1) Successors and predecessors threshold: Same as Harris's original approach.

(2) Peaks versus plateaus: Every peak or plateau is considered as a segmentation position.

(3) Complete word: A prefix is considered as a segment if and only if it matches a word in the corpus.

(4) Entropy: Let $w = a_0 \dots a_{n-1}$ be a word of n letters (a_i). Let w_i be the prefix $a_0 \dots a_{i-1}$ and $cnt(w_i)$ the number of words in the corpus sharing the prefix w_i. The entropy at a position i is as follows

$$H_{w_i} = \sum_{l \in \mathcal{A}} -\frac{cnt(w_i.l)}{cnt(w_i)} \times \log \frac{cnt(w_i.l)}{cnt(w_i)}$$

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

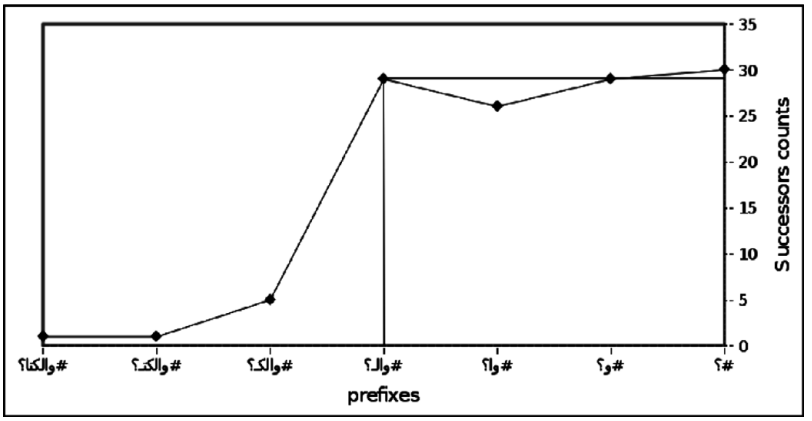


Fig. 2. Successors counts variation for the word 'والكتاب'.

where l is a letter over an alphabet $\mathscr{A}$ and '.' the string concatenation. The position at which the entropy exceeds a threshold is considered as cutoff point.

Xu and Croft (1998) proposed a co-occurrence-based measure to build a stemmer or refine an existing one. Their basic hypothesis is *the word forms that should be conflated for a given corpus will co-occur in documents from that corpus.* The same approach has been tested on Arabic (Larkey, Ballesteros and Connell 2002) and results show that the co-occurrence approach alone is not so effective for Arabic as it is for English and Spanish.

Rogati, McCarley and Yang (2003) have used a statistical translation with an English stemmer and a parallel corpus to build an unsupervised Arabic stemmer, although the building itself is supervised. The principle is to build a translation model using a parallel corpus. The translation model allows them to map the English stems carried out by an English stemmer to their Arabic translations. This initial Arabic stemmer is deployed to segment words of the Arabic side of the parallel corpus. These two steps are iterated to improve both the accuracy of the translation model and the Arabic stemmer.

Lee *et al.* (2003) used a manually segmented corpus (suffixes, prefixes dictionary) to build a stemmer. Their stemmer attempts to carry out all possible segmentations following the template prefix*.stem.suffix* then checks them against the manually built corpus.

Creutz and Lagus (2002) proposed two morpheme-like segmentation methods. The first based on Minimum Description Length (MDL) (Grnwald 2007), and the second based on Maximum Likelihood (ML).

Sharma, Kalita and Das (2002) used a set of heuristics to build a dictionary of segmented words for Assamese language. Monson *et al.* (2004) propose an automatic extraction of inflectional classes from a corpus. Claimed to be unsupervised and language independent, the technique was tested for Spanish.

Karagol-Ayan, Doermann and Weinberg (2006) use longest common substring and edit distance to segment into stem-affix words in a limited and eventually noisy dictionary.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

Other unsupervised morphology related approaches have been investigated in the other languages (Goldsmith 2001; Snover, Jarosz and Brent 2002).

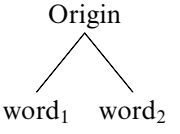
A more simplistic approach, often said light stemming, is to strip out unconditionally all common prefixes and suffixes (Chen and Gey 2002; Larkey Ballesteros and Connell 2007). This seems to perform acceptable results in information retrieval. This may be due to the fact observed and used in (Xu and Croft 1998).

Baroni, Matiasek and Trost (2002) proposed to extract a list of related words from a plain text using two metrics: the edit distance and mutual information, where Schone and Jurafsky (2000) use a combination of distribution and semantic information.

In Section 3, we describe our approach intended to define a metric that assesses how morphologically close are two words to each other. This metric could be simplified to the binary decision: whether two words are morphologically related or not. Section 4 discusses the tests and results.

2 Morphological relatedness

In word variants matching, stemming and radicals extraction are the process of going up to the morphological origins of the two compared variants to check whether these origins match or not.



This approach faces two main challenges:

(1) Which origin is appropriate?
(2) How accurate is the origin retracing?

The first challenge is the known debate on which is the best normal form: stem or root.[1] The second challenge has been paid much more attention by the research community. This resulted in many stemming approaches, where less mature approaches are developed for radicals extraction (de Roeck and Al-Fares 2000; Daya, Roth and Wintner 2008; Boudlal *et al.* 2011).

We claim that another option is to compare directly the two words in order to say how morphologically close they are to each other. In other terms, how likely are they derived from the same morphological origin? With no need to compute that origin, in this section we discuss our approach to measure this closeness.

Let w_1 and w_2 be the two words whose morphological relatedness we would like to know . Let $freq(a)$ be the frequency (count) of the letter $a \in \mathscr{A}$ observed in some corpus. $LCS(w_1, w_2)$ is the longest sequence common to w_1 and w_2 (Hsu and Du 1984; Crochemore, Hancart and Lecroq 2007) and $\overline{LCS(w_1, w_2)}$ all letters of w_1 and

[1] We will say radicals or root to designate the morphological origin from which an Arabic word is derived. In the literature this may be called root, radicals, triliteral root,...

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

w_2 not involved in $LCS(w_1, w_2)$. $\|w\|$ is the size of the word $w \in \mathscr{A}^*$. Finally $w[i]$ is the i^{th} letter of the word w.

The morphological relatedness MR is computed as follows:

$$MR = \sum_{i=1}^{\|LCS(w_1,w_2)\|} \log\left(\frac{1}{freq(LCS(w_1,w_2)[i])}\right) - \sum_{i=1}^{\|\overline{LCS(w_1,w_2)}\|} \log\left(\frac{1}{freq(\overline{LCS(w_1,w_2)}[i])}\right)$$

Example 2.1

Let $w_1 =$'كتاب' ([kɪtɑːb]: a book) and $w_2 =$'كتب' ([kutub]: books).

We have

$$LCS(w_1, w_2) = \text{'ك ت ب'}$$

Then

$$MR(w_1, w_2) = \log\left(\frac{1}{freq(\text{'ك'})}\right) + \log\left(\frac{1}{freq(\text{'ت'})}\right) + \log\left(\frac{1}{freq(\text{'ب'})}\right) - \log\left(\frac{1}{freq(\text{'ا'})}\right)$$

2.1 The intuition beneath

An Arabic word is a sequence of (with respect to our purpose) letters of two types: (1) affixes, and (2) core letters. Affixes are common to all words built following the same pattern as the suffix "ing" is common to English words "*ing". For instance, the letter 'ا' is a common infix to Arabic words built following the pattern "؟ا؟؟" such as:

ب	ا	ت	ك	: Book
ء	ا	ن	ب	: Building
س	ا	ر	غ	: Planting

The idea is that letters in affixes are expected to be of high frequencies since they often participate in words even when they are not of their core structure. However, the core letters are supposed to be of much lower frequencies. Using the formula above:

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

- Lower are the frequencies of the LCS letters, higher is the MR.
- Longer is LCS, higher is the MR.
- Longer is $\overline{LCS}$, lower is the MR.
- Lower are the frequencies of $\overline{LCS}$ letters, lower is the MR.

This reflects the intuition:

- When a LCS letter has a low frequency, this means that most likely the two words share a core letter.
- When a LCS letter has a high frequency, this means that most likely the shared letter is an affix one.

Analogically:

- When a $\overline{LCS}$ letter has a low frequency, this means that most likely a core letter of one word is missing in the second.
- When a $\overline{LCS}$ letter has a high frequency, this means that most likely only an affix letter in the one word is missing in the second.

Hence:

- More core letters are common, more likely the two words are derived from the same morphological origin.
- More core letters are uncommon, more likely the two words are not derived from the same origin.

2.2 MR interpretation

One of the objectives is to use the MR to find the word classes, where words are of the same class if and only if they are derived from the same morphological origin. Let w_0 be an Arabic word and $S = w_1, w_2, \ldots, w_n$ a set of Arabic words. We would like to extract five words in S that are variants (i.e., derived from the morphological origin) of w_0 using only the MR measure. We have to:

(1) compute all MRs: $MR(w_0, w_1), MR(w_0, w_2), \ldots, MR(w_0, w_n)$,
(2) sort the MRs in a descending order, and
(3) keep the five words with the five highest MRs.

Of course, there may be less (eventually none) than five actual variants in S and the steps above will lead to a wrong result. However:

- In a real application such as information retrieval systems, the number of sought variants could be adjusted to fit the required results.
- A threshold for the significant MR can be defined.

2.3 Complexity

$LCS(w_1, w_2)$ can be computed in $O(\|w_1\| \times \|w_2\|)$ time and $O(\min(\|w_1\|, \|w_2\|))$ space (Crochemore *et al.* 2007). $freq$ is computed over a corpus only once and requires a table of $O(\#\mathscr{A})$ space where $\#\mathscr{A}$ is the size of the alphabet. The

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

access to a *freq*'s entry requires a constant time. Therefore, the time and space complexities of the morphological relatedness computation depend only on the time and space complexities of the computation of $LCS(w_1, w_2)$, which are respectively $O(\|w_1\| \times \|w_2\|)$ time and $O(\min(\|w_1\|, \|w_2\|))$ space.

3 Tests and results

Our investigation targeted traditional formal Arabic for the following reasons:

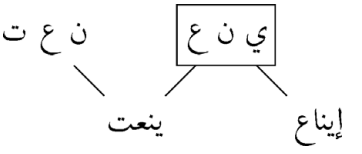
(1) Traditional Arabic has many differences in terms of vocabulary as well as grammar with the modern Arabic in use.
(2) Many collections are in pure traditional Arabic.
(3) To the best of our knowledge, no published work targets explicitly the old traditional Arabic.
(4) Basically, generalizing our approach to a lesser restrictive vocabulary such as modern Arabic should be straight.

Performance of an information retrieval (IR) system involves factors other than the normalization approach. It should be then arguable to evaluate the measure independently of any information retrieval system.

We consider the binary relevance in which two words w_1 and w_2 are relevant to each other if they share at least one root. For instance, the word 'ينعت' has at least two meanings since it has at least two roots:

Word	Transliteration	Meaning	Root
ينعت	jaʃatu	is showing	'ن ع ت'
	jaʃaʃat	has ripened	ي ن ع

The word 'إيناع' ([ʔiːnaːʃun]: Ripening) has the root 'ي ن ع'. 'ينعت' and 'إيناع' are then relevant to each other since they share the root 'ي ن ع'.



3.1 The test dataset

We have extracted the letters' frequencies (*freq*) from a corpus exceeding 500k distinct words dumped from an old Arabic dictionary (فيروزأبادي 1965). We then mapped more than 200k words to their roots using ElixirFM (Smrž 2007) morphological analyzer, which is based on Tim Buckwalter's morphological analyzer (Buckwalter 2004).

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

For every word in a sample of 80,629 words randomly chosen among the 200 k words, we have:

(1) Extracted the set of all related words $R_{root}(w)$ in the 200k corpus where:

$$R_{root}(w) = \{w_i \in \text{Corpus} : rel(w_i, w) = 1\}$$

Where:

$$rel(w_i, w) = \begin{cases} 1 & \text{if } roots(w) \cap roots(w_i) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

where $roots(w)$ is the set of w's roots.

(2) Extracted the set $\{w_1, \ldots, w_k\}$ of k words having respectively the k highest MR scores with $k = \#roots(w)$ (cardinality of $roots(w)$).

(3) Measured the Normalized Discounted Cumulative Gain (NDCG) (Sakai 2007) and the R-precision (Buckley and Voorhees 2000; Aslam and Yilmaz 2005). In our case :

$$NDCG(w) = \frac{DCG(w)}{IDCG(w)}$$

Where

$$DCG(w) = rel(w_1, w) + \sum_{i=2}^{k} \frac{rel(w_i, w)}{\log_2(i)}$$

$$IDCG(w) = 1 + \sum_{i=2}^{k} \frac{1}{\log_2(i)}$$

Intuitively, the NDCG reflects how relevant and ordered are the results. The prefect NDCG is 1, which occurs when all the results are correctly ordered and all of them relevant. In our case, this occurs when all results share at least one root with w since we adopted the binary relevance. When a non-relevant result is carried out, the NDCG decreases depending on the ranking of that non-relevant result. If the non-relevant result is carried out first, the NDCG will be lower than when the non-relevant result is carried out at the second or n^{th} position. Higher is n, higher is the NDCG. However, R-precision shows among the results how many are actually relevant, and is calculated as follows:

$$R - \text{precision} = \frac{\sum_{i=1}^{k} rel(w_i, w)}{k}$$

3.2 Collation

In order to avoid mismatches induced by variations in letter forms, we introduce the following normalization.

(1) 'آ' has been replaced by 'ءا'

(2) 'أ', 'إ', 'ؤ', and 'ئ' by 'ء'

(3) finally, 'ة' by 'ت'

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

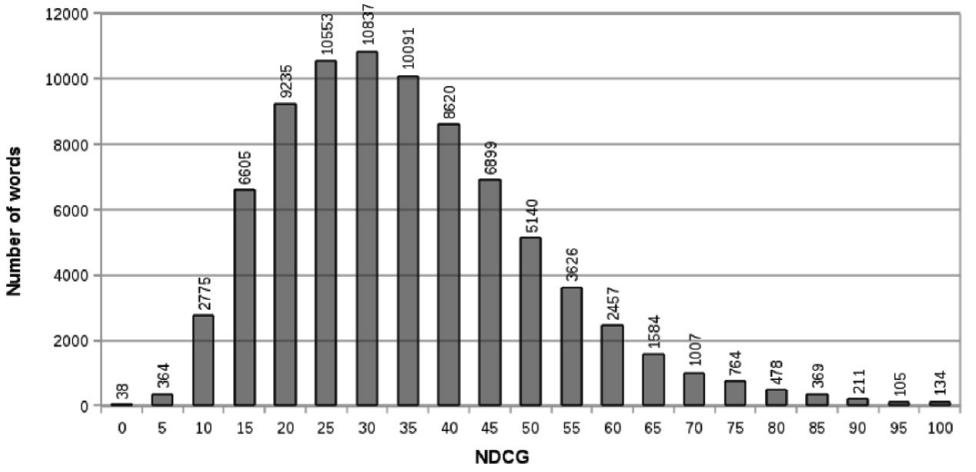


Fig. 3. NDCGs distribution.

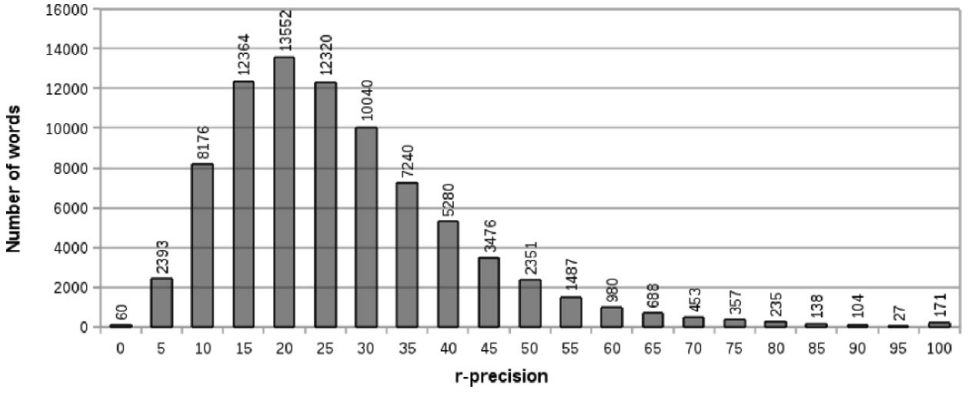


Fig. 4. R-precisions distribution.

A subset of the collation rules given above has been tested (Khorsi 2012) and the results show its positive impact.

3.3 Results and discussion

The global averages of NDCG and R-precision reached respectively 34.89% and 26.88%. Figures 3 and 4 show the NDCG and R-precision distribution over the test sample.

Every bar gathers the words with regard to their rounded NDCGs and R-precisions. For instance, 9,235 words have their NDCGs around 20% and 13,552 have their R-precision around 20%. Most of the results do not reach 50% of either the NDCG or the R-precision. Yet a significant portion exceeds the 50% in both measures.

In the paragraphs below, we examine the NDCG and R-precision values variations against various factors to see whether a clear influence of any could be distinguished

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

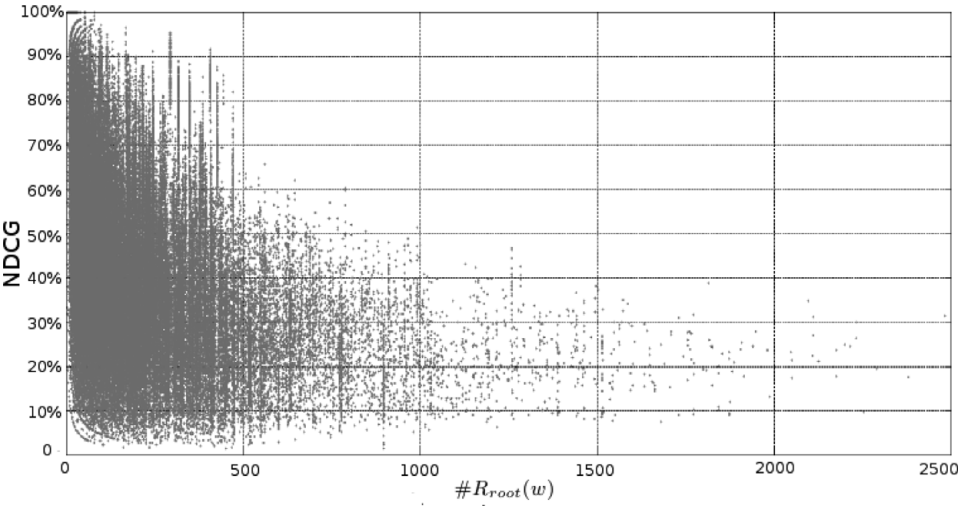


Fig. 5. NDCGs variation over $\#R_{root}(w)$.

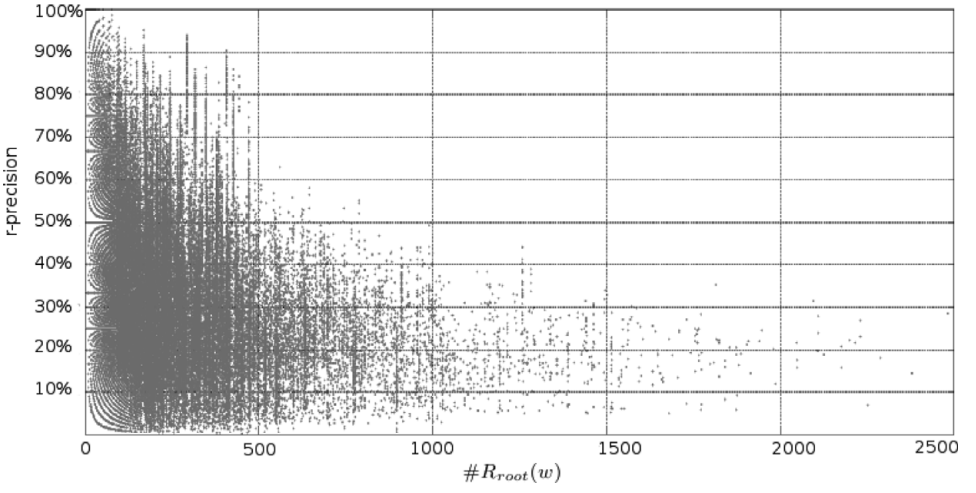


Fig. 6. R-precisions variation over $\#R_{root}(w)$.

and thereby to see what makes the NDCG and R-precision increase and what makes them decrease. One suspected factor is the relative length of affixes. Intuitively, when the prefix and/or the suffix is too long, this will have two main effects:

(1) A variant which has not that long affix will have its MR low since all letters of the affixes will be discounted.
(2) A non-variant word which has that long affix will have its MR high since all common letters are counted in the MR and core letters which are not common will not make the MR decrease enough.

Let a word $w = u{\cdot}w_0{\cdot}v$ and its variant $w' = u'{\cdot}w_0{\cdot}v'$. Let another word $x = u{\cdot}x_0 \cdot v$ where '$\cdot$' is the string concatenation. w and w' share w_0 where they differ in $u.v.u'{\cdot}v'$. However, w and x share $u{\cdot}v$ and differ in $w_0 \cdot x_0$.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

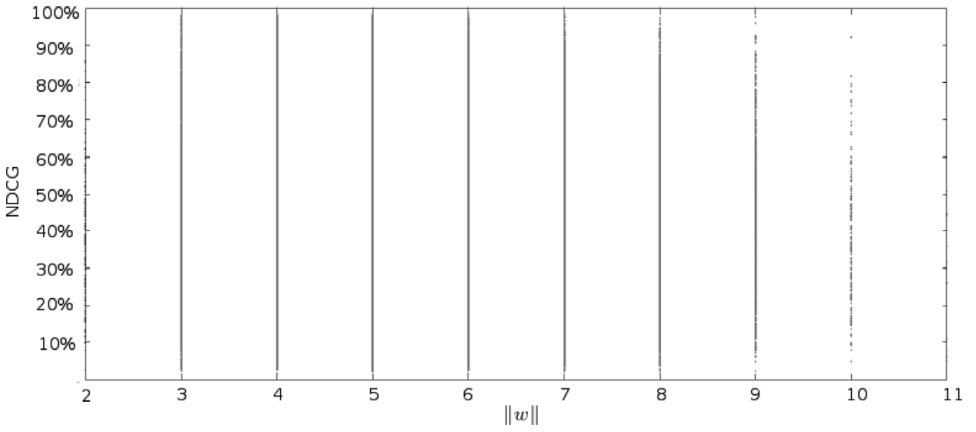


Fig. 7. NDCGs variation over $\|w\|$.

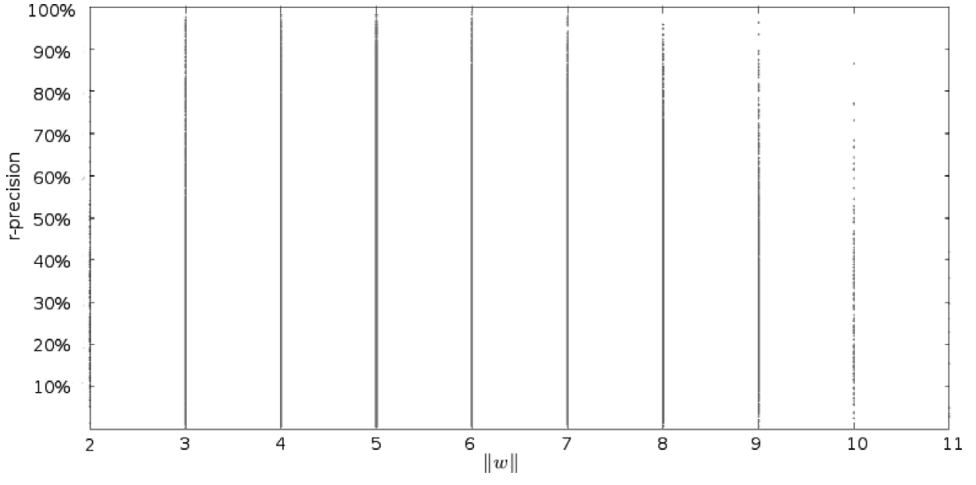


Fig. 8. R-precisions variation over $\|w\|$.

Figures 5 and 6 show the variations against the numbers of related words $\#R_{root}(w)$.

Both give the impression that words with high form variability tend to have relatively smaller portion of variants covered by the MR measure, which is terminologically said low recall. This is due to the fact that a high number of variants means occurrence of a high number of inflectional and mutating letters such as long vowels. Even with high frequencies, their abundance reduces the MR and makes words not actually relevant but with lesser inflectional letters jump up in the ranking. On the counter part, the only words reaching a perfect ranking 100% are those with low variability.

For sizes of words (see Figures 7 and 8).

Unlike Harris-based approaches, which are affected by the length of words and affixes, our's does not seem to show any influence of the word length. Notice that

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

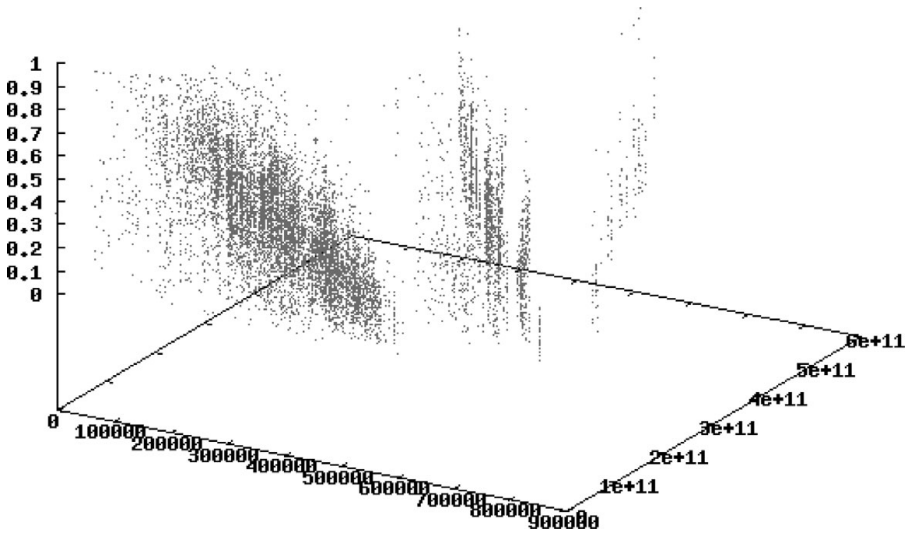


Fig. 9. Impact of the mean and variance of frequencies of *root*(*w*)'s letters on NDCG.

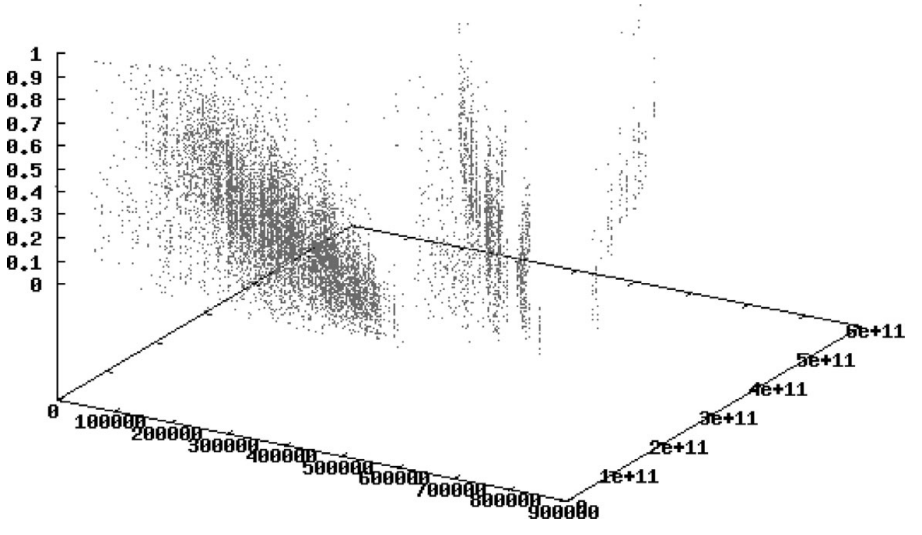


Fig. 10. Impact of the mean and variance of frequencies of *root*(*w*)'s letters on R-precision.

the longest word in our corpus is that of eleven letters, the smallest is that of two, and the length average is 5.38.

Figures 9 and 10 show the impact of the mean and variance of the frequencies of letters in *root*(*w*).

As expected, in both, the main part of results shows that the more letters of the roots are of lower frequencies, the higher are the NDCG and R-precision. Indeed, the lower are the frequencies of the root letters, the better is the word, and its variants distinguishable.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

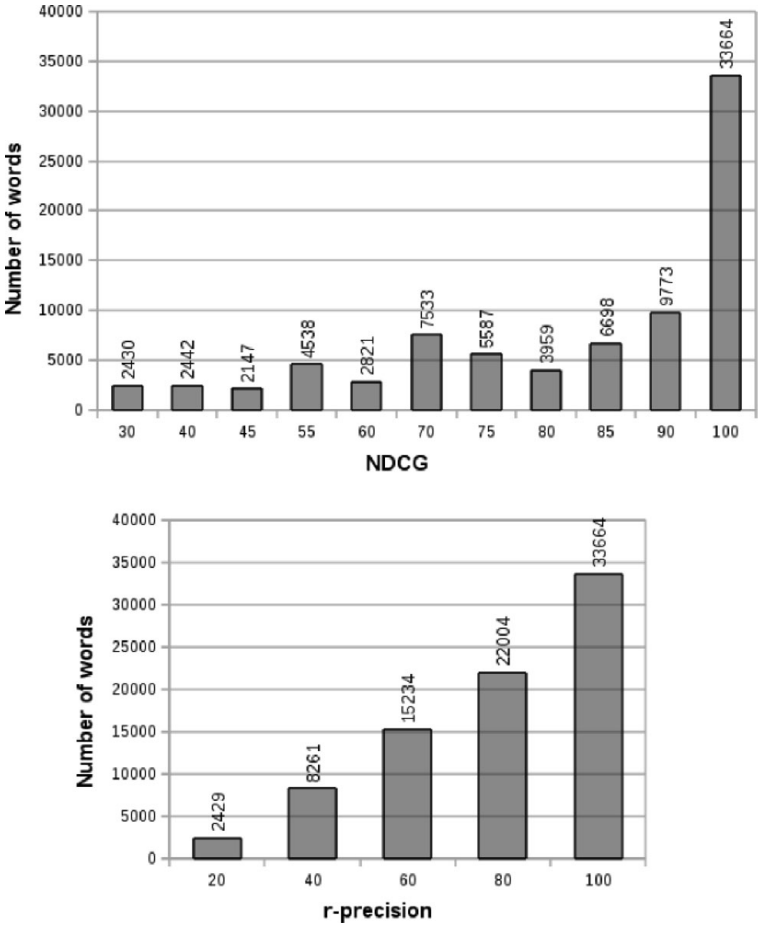


Fig. 11. NDCG and r-precision for the top five variants.

3.4 Results for a fixed number of top variants

The analysis above was on the overall results. In the paragraphs below, we examine the performance of our approach in extracting a fixed number of variants.

For five top variants, the NDCG and R-precision reached respectively 82.76% and 78.68% (see Figure 11).

This clearly shows a good effectiveness in the extraction of a fair number of variants. Analogically, Figures 12 and 13 show respectively the NDCG and R-precision for different numbers of variants.

For a wider view, Figure 14 shows how NDCG and R-precision are varying over the number of the extracted variants.

4 Conclusion

We discussed results of the use of an unsupervised evaluation of words' morphological relatedness. This measure is intended to assess how likely two words are

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

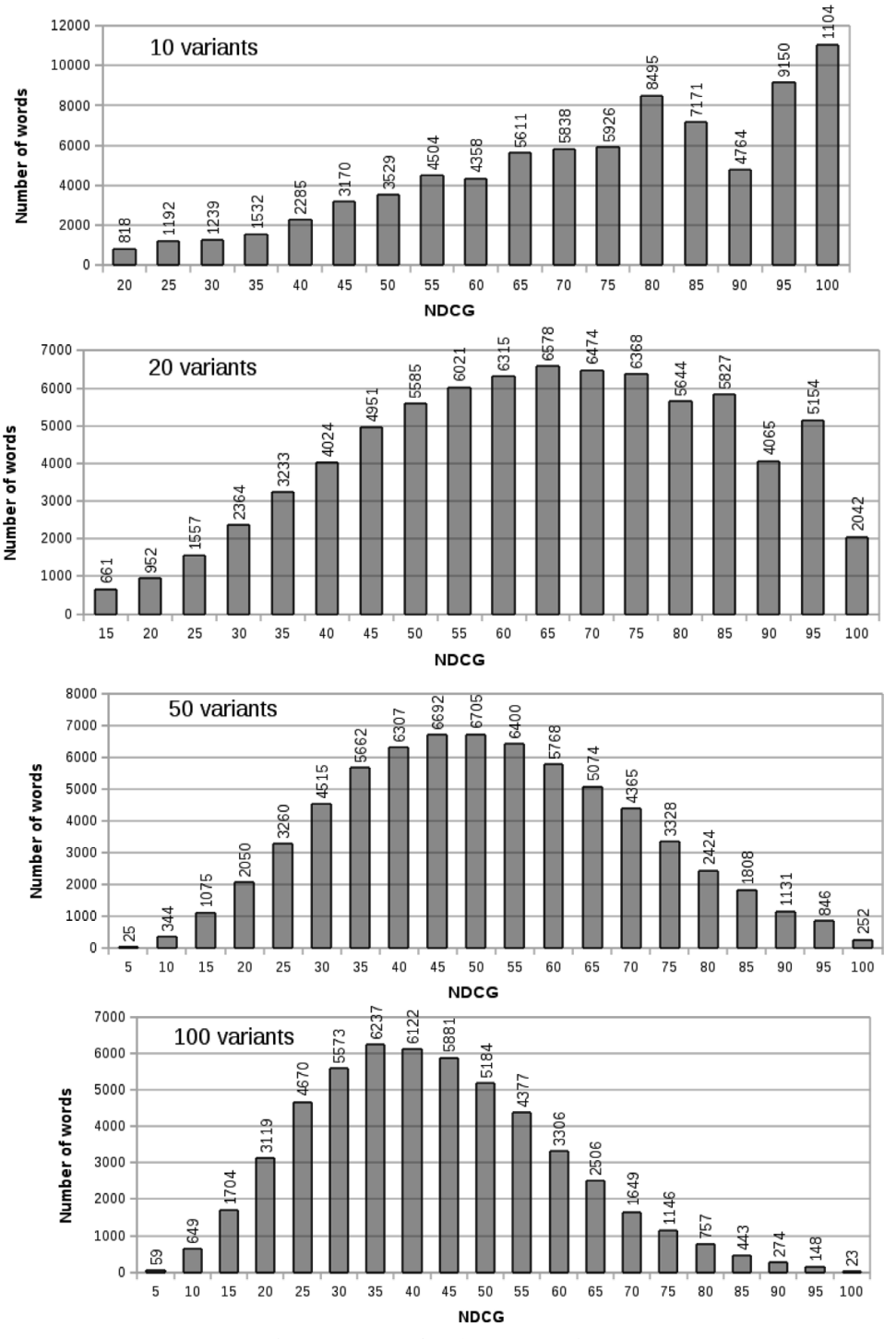


Fig. 12. NDCG for the n top variants.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

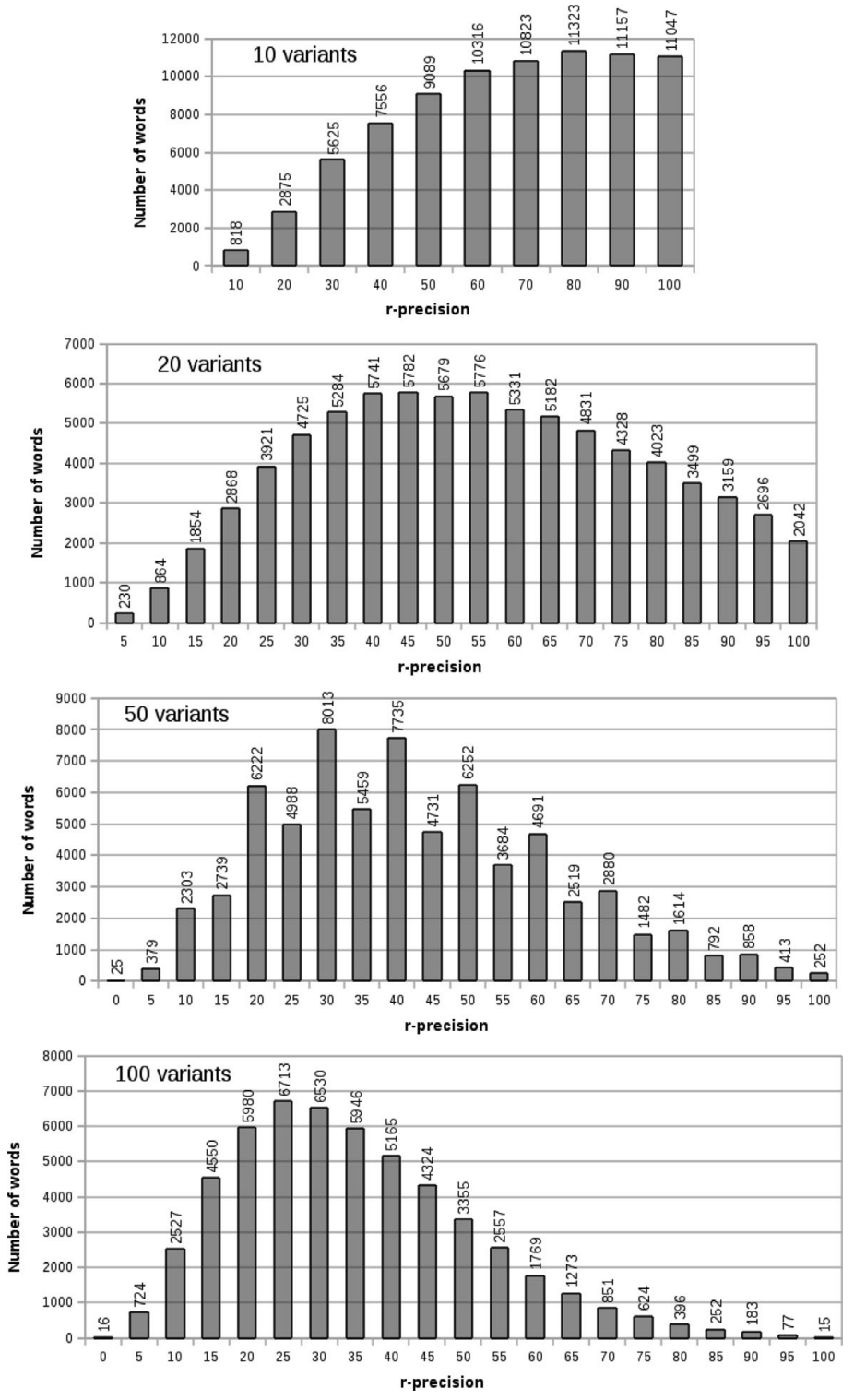


Fig. 13. R-precision for the *n* top variants.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

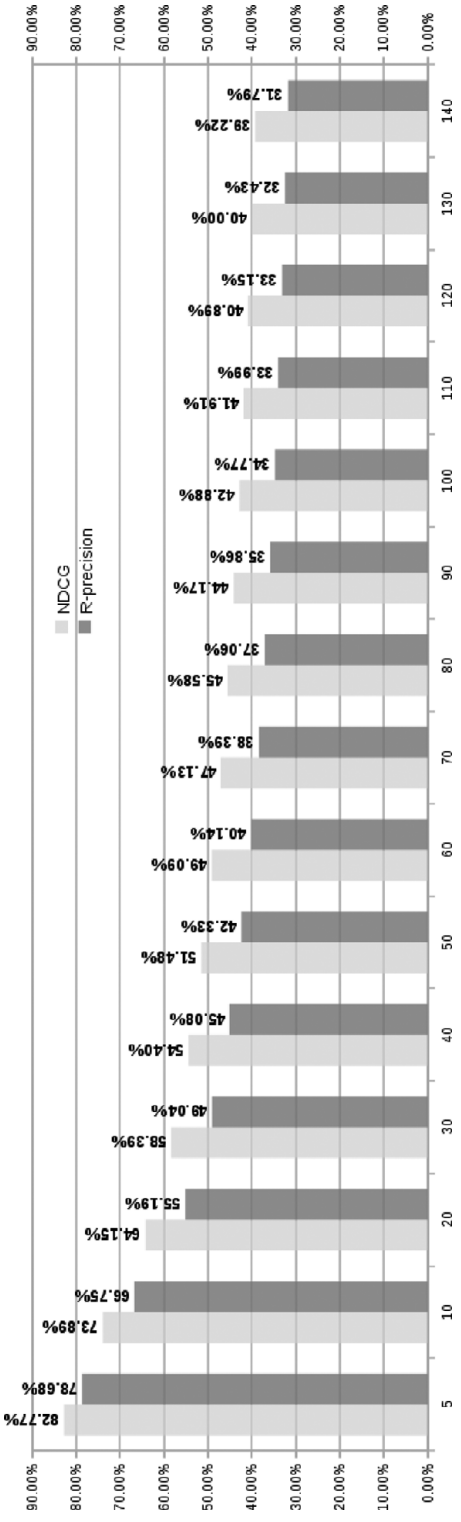


Fig. 14. NDCG and r-precision averages for different sizes of top variants.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

derived from the same morphological origin, which in our study are the radicals. This measure may be viewed as an alternative to radicals extraction and stemming. The main features of our approach are:

- Its computation is entirely unsupervised and no linguistic knowledge is required.
- Its computation is fast and fit requirements of real applications such as information retrieval systems.
- Its time and space complexity depend only on the sizes of the two compared words.
- The comparison does not need the computation of the normal form (i.e., origin, e.g., stem, root). In the traditional approaches, both the normal forms finding and comparison are empirical, which increases the risk of error.

Results show that when the letters of the root have high frequencies, the measure is not able to effectively discriminate the variants. This appeals to think about adding a weighting factor to the measure. One option could be a weight that is refined iteratively. This may make the computation greedier in time. Another improvement is to extend the frequencies to n-grams. When the core letters have high frequencies, their combination with the other core letters will decrease the problematic high frequency and perform better discrimination. On the counter part, it will complicate the choice of the common part used to compute the relatedness. Recall that we have used here the LCS as a common part. Moreover, the use of the n-grams will also decrease the frequencies of the non-core letters making them a confusing factor. Take the example of two unrelated words but sharing the same long prefix such as 'وال' ([wel]: and the). The frequency of the whole prefix is much lower than its separated letters. Another potential improvement is to introduce an unsupervised stemming before the computation of the MR. This will avoid the problem of long affixes highlighted above. Still, our measure shows a good effectiveness on the extraction of a fair number of variants.

References

Aslam, J. A., and Yilmaz, E. 2005. A geometric interpretation and analysis of R-precision. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken (eds.), *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, pp. 664–71, Bremen, Germany.

Baroni, M., Matiasek, J., and Trost, H. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning MPL '02*, pp. 48–57, Stroudsburg, PA.

Boudlal, A., Belahbib, R., Lakhouaja, A., Mazroui, A., Meziane, A., and Bebah, M. 2011. A Markovian approach for Arabic root extraction. *International Arab Journal of Information Technology* **8**(1): 91–8.

Buckley, C., and Voorhees, E. M. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, Athens, Greece.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

Buckwalter, T. 2004. Issues in Arabic orthography and morphology analysis. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, pp. 31–34, Stroudsburg, PA.

Chen, A., and Gey, F. 2002. Building an Arabic stemmer for information retrieval. In *Proceedings of TREC 2002*, pp. 631–39, Gaithersburg, MD.

Creutz, M., & Lagus, K. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, MPL '02, pp. 21–30, Stroudsburg, PA.

Crochemore, M., Hancart, C., & Lecroq, T. 2007. *Algorithms on Strings*. 1st ed. Cambridge University Press.

Daya, E., Roth, D., and Wintner, S. 2008. Identifying Semitic roots: machine learning with linguistic constraints. *Computational Linguistics* **34**(3): 429–48.

de Roeck, A. N., and Al-Fares, W. 2000. A morphologically sensitive clustering algorithm for identifying Arabic roots. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 199–206, Hong Kong.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* **27**(2): 153–98.

Grnwald, P. D. 2007. *The minimum description length principle*. Cambridge, MA: MIT Press.

Hafer, M. A., & Weiss, S. F. 1974. *Word Segmentation by Letter Successor Varieties*. Amsterdam: Elsevier.

Harris, Z. S. 1955. From phoneme to morpheme. *Language* **31**(2): 190–222.

Hsu, W. J., and Du, M. W. 1984. New algorithms for the LCS problem. *Journal of Computer and System Sciences* **29**: 133–52.

IPA (International Phonetic Association). 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, England, UK: Cambridge University Press.

Jurafsky, D., & Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Karagol-Ayan, B., Doermann, D., and Weinberg, A. 2006. Morphology induction from limited noisy data using approximate string matching. In *Proceedings of the 8th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pp. 60–8, New York, USA.

Khorsi, A. 2012. Effective unsupervised Arabic word stemming: towards an unsupervised radicals extraction. *IAJIT* **9**(6).

Larkey, L. S., Ballesteros, L., and Connell, M. E. 2002. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–82, Tampere, Finland.

Larkey, L., Ballesteros, L., and Connell, M. 2007. Light stemming for Arabic information retrieval. In N. Ide, J. Vronis, H. Baayen, K. W. Church, J. Klavans, D. T. Barnard, D. Tufis, J. Llisterri, S. Johansson, J. Mariani, A. Soudi, A. van den Bosch, & G. Neumann (eds.), *Arabic Computational Morphology*, pp. 221–43. Text, Speech and Language Technology, vol. 38. Dordrecht, The Netherlands: Springer.

Lee, Y.-S., Papineni, K., Roukos, S., Emam, O., and Hassan, H. 2003. Language model based Arabic word segmentation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pp. 399–406, Sapporo, Japan.

Menn, L. 1995. *Non-Fluent Aphasia in a Multilingual World*. Amsterdam: John Benjamins.

Monson, C., Lavie, A., Carbonell, J., & Levin, L. 2004. Unsupervised induction of natural language morphology inflection classes. In *Proceedings of the 7th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pp. 52–61, Barcelona, Spain.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press

Rogati, M., McCarley, S., & Yang, Y. 2003. Unsupervised learning of Arabic stemming using a parallel corpus. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pp. 391–98, Sapporo, Japan.

Sakai, T. 2007. On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management* **43**(2): 531–48.

Schone, P., and Jurafsky, D. 2000. Knowledge-free induction of morphology using latent semantic analysis. In C. Cardie, W. Daelemans, C. Nédellec, and E. T. K. Sang (eds.), *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, pp. 67–72, Lisbon, Portugal.

Sharma, U., Kalita, J., and Das, R. 2002. Unsupervised learning of morphology for building lexicon for a highly inflectional language. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pp. 1–10, Philadelphia, USA.

Smrž, O. 2007. Elixirfm: implementation of functional Arabic morphology. In *Semitic '07: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages*, pp. 1–8, Morristown, NJ, USA.

Snover, M. G., Jarosz, G. E., and Brent, M. R. 2002. Unsupervised learning of morphology using a novel directed search algorithm: taking the first step. *In Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, pp. 11–20, Philadelphia, PA, USA.

Soudi, A., & van den Bosch, A. 2007. *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*. Dordrecht, The Netherlands: Springer.

Sproat, R. W. 1992. *Morphology and Computation*. Cambridge, MA: MIT Press.

Xu, J., and Croft, W. B. 1998. Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems* **16**(1): 61–81.

فيروزأبادي. 1965. تاج العروس من جواهر القاموس، وزارة الارشاد والانباء.

https://doi.org/10.1017/S1351324912000071 Published online by Cambridge University Press