Corpus-based Lexeme Ranking for Morphological Guessers

Krister Linden and Jussi Tuovila

University of Helsinki, Helsinki, Finland

Abstract. Language software applications encounter new words, e.g., acronyms, technical terminology, loan words, names or compounds of such words. To add new words to a morphological lexicon, we need to determine their base form and indicate their inflectional paradigm. A base form and a paradigm define a lexeme. In this article, we evaluate a lexicon-based method augmented with data from a corpus or the internet for generating and ranking lexeme suggestions for new words. As an entry generator often produces numerous suggestions, it is important that the best suggestions be among the first few, otherwise it may become more efficient to create the entries by hand. By generating lexeme suggestions with an entry generator and then further generating some key word forms for the lexemes, we can find support for the lexemes in a corpus. Our ranking methods have 56-79 % average precision and 78-89 % recall among the top 6 candidates, i.e. an F-score of 65-84 %, indicating that the first correct entry suggestion is on the average found as the second or third candidate. The corpus-based ranking methods were found to be significant in practice as they save time for the lexicographer by increasing recall with 7-8 % among the top candidates.

1 Introduction

New words are constantly finding their way into daily language use. This is particularly prominent in rapidly developing domains such as biomedicine and technology. The new words are typically acronyms, technical terminology, loan words, names or compounds of such words. They are likely to be unknown by most hand-made morphological analyzers. In many applications, hand-made guessers are used for covering the low-frequency vocabulary or the strings are simply added as such.

Mikheev [13, 14] pointed out that words unknown to the lexicon present a substantial problem for part-of-speech tagging, and he presented a very effective supervised method for inducing English guessers from a lexicon and an independent training corpus. Oflazer & al. [15] presented an interactive method for learning morphologies and pointed out that an important issue in the wholesale acquisition of open-class items is that of determining which paradigm a given citation form belongs to.

Recently, unsupervised acquisition of morphologies from scratch has been studied as a general problem of morphology induction in order to automate the morphology building procedure. For overviews, see Wicentowski [18] and

Goldsmith [3]. If we do not need a full analysis, but only wish to segment the words into morph-like units, we can use segmentation methods like Morfessor [1]. For a comparison of some recent successful segmentation methods, see the Morpho Challenge [8].

Although unsupervised methods have some advantages for less-studied languages, for the well-established languages, we have access to fair amounts of lexical training material in the form of analyses in the context of more frequent words. Especially for Germanic and Finno-Ugric languages, new words tend to be compounds of acronyms and loan words with existing words. For these languages, there are already large-vocabulary descriptions available. In English, compound words are written separately or the junction is indicated with a hyphen, but in other Germanic languages and in the Finno-Ugric languages, there is usually no word boundary indicator within a compound word. It has previously been demonstrated by Lindén [9] that already training sets as small as 5000 word forms and their manually determined base forms will give a reasonable result for guessing base forms of new words by analogy. The experiments were performed on a set of languages representing different language families, i.e. English, Finnish, Swedish and Swahili.

In addition, there are a host of large but shallow hand-made morphological descriptions available, e.g., the Ispell collection of dictionaries [7] for spell-checking purposes, and many well-documented morphological analyzers are commercially available, e.g. [12]. It has also been demonstrated by Lindén [10] that there is a simple but efficient way to derive an entry generator from a full-scale morphological analyzer implemented as a finite-state transducer. Such an entry generator can be used as a baseline for more advanced entry guessing methods.

Using the entry generator developed by Lindén [10], we can generate lexeme candidates, i.e. base form and paradigm combinations, for new words with the entry generator and then further generate key word forms for the lexeme candidates. Using these lexeme candidates with key word forms, a person with native skills can select the correct entry for a new word. With this method, we encoded a set of words based on an open source dictionary project with a different encoding scheme than ours[1]. We selected all the words that were unknown to our lexicon and used the entry generator to encode the new words according to the guidelines of *The Research Institute for the Languages of Finland* [6] used in our lexicon. The reclassification took approximately 20 hours of work during which a list of 11 026 new entries was created. As the words had been categorized, we were able to take advantage of the existing categories to guide the process, but a number of systematic mismatches and ambiguities between the two encoding schemes exist. The work was a considerable speed-up compared to hand-coding the words from scratch, but manually disambiguating between lexeme candidates is still tedious work, and it motivated the current research to find additional methods for speeding up the encoding task.

In this article, we propose and evaluate new methods for *ranking lexeme suggestions for a word form* of a new word by generating lexeme candidates,

[1] http://joukahainen.puimula.org/

i.e. base form and paradigm combinations, with an entry generator and then further generating key word forms[2] for the lexeme candidates in order to *find support for the lexemes in a corpus* to weed out irrelevant lexeme suggestions. In Sect. 2, we outline the directly related previous work. In Sect. 3, we describe the new methods. In Sect. 4, we present the training and test data. In Sect. 5, we evaluate the model. In Sect. 6, we discuss the method and the test results in light of the existing literature and some similar methods.

2 Lexicon-based Entry Generator

To create entries for a morphological analyzer from previously unseen words and word forms, we need an entry generator. Ideally, we can use information that is already available in some existing morphological description to encode new entries in a similar fashion. Below, we briefly outline a general method for creating lexicon-based entry generators that was introduced by Lindén [10]. In his article, Lindén demonstrates that the method works well for English, Finnish and Swedish.

Assume that we have a finite-state transducer lexicon T which relates base forms, $b(w)$, to word forms, w . Let w belong to the input language L_I and $b(w)$ to the output language L_O of the transducer lexicon T . Our goal is to create an entry generator for word forms that are unknown to the lexicon, i.e. we wish to provide the most likely base forms $b(u)$ for an unknown input word u. In order to create an entry generator, we first define the left quotient and the weighted universal language with regard to a lexical transducer. For a general introduction to automata theory and weighted transducers, see e.g. [17].

We can regard the left quotient as the set of postfixes of L_1 that complete words from L_2 such that the resulting word is in L_1 . If L_1 and L_2 are formal languages, the left quotient of L_1 with regard to L_2 is the language consisting of strings w such that xw is in L_1 for some string x in L_2 . Formally, we write the left quotient as in (1).

$$L_1 \backslash L_2 = \{a | \exists x((x \in L_2)(xa \in L_1))\} \tag{1}$$

If L is a formal language with alphabet Σ, a universal language, U, is a language consisting of strings in Σ^*. The weighted universal language, W, is a language consisting of strings in Σ^* with weights $p(w)$ assigned to each string. For our purposes, we define the weight $p(w)$ to be proportional to the length of w. We define a weighted universal language as in (2).

$$W = \{w | \exists w(w \in \Sigma^*)\} \tag{2}$$

[2] In highly inflecting languages like Finnish, it is not feasible to generate all word forms of a paradigm, as a noun can have more than 2000 word forms and a verb more than 10000 forms. A paradigm can be identified by a small set of inflected forms. This strategy is often used in lexicons intended for language learners to identify or illustrate verb paradigms for irregular verbs, e.g. in Romance or Germanic languages.

with weights $p(w) = C|w|$, where C is a constant.

A finite-state transducer lexicon, T, is a formal language relating the input language L_I to the output language L_O. The pair alphabet of T is the set of input and output symbol pairs related by T. An identity pair relates a symbol to itself.

We create an entry generator, G, for the lexicon T by constructing the weighted universal language W for identity pairs based on the alphabet of L_1 concatenating it with the left quotient of T with regard to the universal language U of the pair alphabet of T as shown in (3).

$$G(T) = WT\backslash U \tag{3}$$

Lindén [10] proves that it is always possible to create an entry generator, $G(T) = WT\backslash U$, from a weighted lexical transducer T.

The model is general and requires no information in addition to the weighted lexicon from which the entry generator is derived. Therefore Lindén suggests that it be used as a baseline for other entry generator methods. For a sample output from the entry generator, see Table 1.

3 Corpus-based Lexeme Ranking

Assume that we have a morphological entry generator that generates a set of base form and paradigm combinations for out-of-vocabulary word forms. Each base form and paradigm combination defines a lexeme. In order to automatically score the lexemes suggested by the entry generator, we generate key word forms of the lexemes and look for the word forms in a corpus. Generally a lexeme whose key word forms are well-attested, i.e. many forms are in use and each form is used repeatedly, is more likely to be correct than a lexeme whose key word forms cannot be found or have only a few documented instances. Rare forms may even be spelling errors. By scoring all the lexemes provided by the entry generator, we can order the lexemes in descending order of support.

We have an unknown word form, w, for which we generate a set of lexeme candidates $U = \{l_1, l_2, l_3, \ldots l_n\}$. Each lexeme candidate, l_i, is defined by its set of word forms, from which we choose a set of key word forms, $K_i = \{k_1, k_2, k_3, \ldots k_m\}$, for scoring support of the lexeme.

We define a method for scoring the possible lexemes of an unknown word form by defining the probability, $P(l_i|w)$, of a lexeme, l_i, when given an unknown word form, w. Since we cannot directly estimate the conditional probability of a lexeme with regard to a word form from a corpus of running text, we use Bayes' rule as in (4) to reformulate the conditional probability.

$$P(l_i|w) = \frac{P(l_i, w)}{P(w)} = \frac{P(w|l_i) * P(l_i)}{P(w)} \tag{4}$$

The most likely lexeme l is provided by (4) by finding the l_i which maximizes the equation, in which $P(w)$ can be regarded as a constant. We then get (5) for l.

$$l = \arg\max_{l_i} P(w|l_i) * P(l_i) \tag{5}$$

We have $P(w|l_i)$ which is the probability of the original word form for a lexeme candidate, i.e. the probability that w is an inflected form of a lexeme candidate l_i, and the probability $P(l_i)$ of the lexeme in a large corpus. As a lexeme is defined by its set of word forms, the probability of a lexeme in a corpus is the sum of the probabilities of its word forms in the corpus. We simplify the equation by assuming that the key word forms of a lexeme are sufficient to estimate the probability of the lexeme with the remaining word forms contributing a negligible constant addition, i.e. for a highly inflecting language, the key word forms should be chosen so that they represent a significant portion of the probability mass of the word forms of the lexeme. We sum over the key word forms and get (6),

$$l = \arg\max_{l_i} \sum_{k \in K_i} P(w|l_i) * P(k, l_i), \tag{6}$$

where $P(k, l_i)$ is the probability that a word form k belongs to lexeme l_i. To further simplify the equation, we assume that the conditional probability of the original word form in a suggested lexeme $P(w|l_i)$ is constant for our purposes, as no lexemes are suggested in which w could not appear as some inflected form, even if it may not be among the key word forms[3]. As a consequence of the assumption, the most likely lexeme l only depends on the innermost term of our equation, which further simplifies to (7).

$$l = \arg\max_{l_i} \sum_{k \in K_i} P(k, l_i), \tag{7}$$

To find the most likely lexeme, l, it is necessary to estimate the joint probability $P(k, l_i)$ that a key word form k co-occurs with lexeme l_i in a corpus.

3.1 Estimating Lexeme Likelihoods

In order to determine the likelihood that a word form co-occurs with a lexeme, we will look at three different methods for estimating this likelihood from a corpus. All three methods essentially regard the lexemes as small documents and the intention is to rank the documents, i.e. the lexemes, by their support in the corpus.

[3] The even distribution of word forms in a lexeme is an oversimplification seemingly contradicting our previous assumption about key word forms representing the core probability mass. For consistency, we should have exploited the fact that a word form w, which is not among the key word forms, is relatively infrequent in the suggested lexeme by giving the lexeme a lower probability. In practice, we could have taken it into account, e.g. by filtering out lexeme suggestions that did not contain the original word form w among the key word forms effectively giving such lexeme suggestions 0 probability.

Key Word Indicator. The most basic method assumes that all key word forms are equally likely to appear in a lexeme. We define an indicator, $I(k)$, which is 1 or 0 depending on whether the word form k appears in the corpus or not. We call the basic method the *key word indicator scoring* and defined it in (8).

$$P(k, l_i) = \frac{I(k)}{|K_i|}, \tag{8}$$

where $|K_i|$ is the number of key word forms that we investigate[4] for the lexeme l_i. If we always considered the same number of word forms for each lexeme, $|K_i|$ could be ignored. For some languages, it may be possible to look at all the word forms of a lexeme, but for some highly inflecting languages it is practical to use only a few key word forms for each lexeme. The number of key word forms may depend on the paradigm of the lexeme or even on the competing lexeme candidates in which case $|K_i|$ is needed as a normalizing factor.

Key Word Frequency. In order to better take into account the fact that a frequent word form is more significant for a lexeme than an infrequent one, which may even be a spelling error, we also consider the frequency, $F(k)$, of a word form k. However, it is unlikely that the importance of a word form is directly proportional to the frequency, so we consider the logarithm of the frequency. In addition, we need to smooth the frequency function by adding one, which creates a frequency scoring that mimics the indicator function for zero frequency key words, where $log(1) = 0$ and grows monotonically for larger frequencies. We thereby get a scoring method defined in (9) that is a variation of the term frequency for documents in information retrieval. We call this method the *key word frequency scoring.*

$$P(k, l_i) = \frac{\log(F(k))}{|K_i| * C}, \tag{9}$$

where C is a normalizing constant proportional to the logarithm of the number of tokens in the corpus. The constant has no effect on the ranking, but it serves to normalize the scoring into a probability distribution.

Key Word Frequency with Inverse Lexeme Frequency. A word form k may simultaneously belong to the set of word forms K_i of several candidate lexemes l_i. To take into account the distinctiveness of each key word form, we calculate a score similar to the inverse document frequency in information retrieval. The *inverse lexeme frequency*, $ilf(k)$, is equal to the logarithm of the number of lexeme candidates, n, divided by the number of candidates $|k \in K_i|$

[4] The connection with the joint probability $P(k, l_i)$ is not obvious in the formula, however, the same key word form may affect two lexemes differently, e.g. the English word form *works* can be seen as one out of four forms of a verb but as one out of two forms of a noun contributing a different probability mass to each of the lexemes.

in which k is a key word form. The distinctiveness score or the *inverse lexeme frequency* of a key word form is defined in (10).

$$ilf(k) = \log \frac{n}{|k \in K_i|}, \tag{10}$$

The key word frequency method is scaled with the distinctiveness score to yield (10). We call this method the *key word frequency with inverse lexeme frequency scoring*.

$$P(k, l_i) = \frac{\log(F(k))}{|K_i| * C} * ilf(k), \tag{11}$$

The scoring methods can be used with any data that reflects the occurrences of key words. Although we refer to the source of word frequency data as a corpus, the method can be used with other data sources as well. As is described in Sect. 5, we have successfully tested the methods using both corpus material and page frequencies returned by a web search engine. In theory, the scoring methods should work with any data source that reflects the occurrence of words in language use.

4 Training and Test Data

To test our methods for corpus-based ranking of lexemes generated by a lexical entry generator, we use the entry generator for Finnish created by Lindén [10] and implemented with the Helsinki Finite-State Technology tools [4]. In 4.1, we briefly describe the lexical resources used for the finite-state transducer lexicon, which was converted into an entry generator.

Words unknown to the lexicon were drawn from a language-specific text collection. The correct entries for a sample of the unknown words were manually determined. In 4.2, we describe the text collections and, in 4.3, the samples used as test data. In 4.4, we describe the evaluation method and characterize the baseline.

4.1 Lexical Data for the Transducer Lexicon and Entry Generator

Lexical descriptions relate look-up words to other words and indicate the relation between them. A morphological finite-state transducer lexicon relates a word in dictionary form to all of its inflected forms. For an introduction, see e.g. [5].

Our current Finnish morphological analyzer was created by [16] based on the Finnish word list *Kotimaisten kielten tutkimuskeskuksen nykysuomen sanalista* [6], which contains 94 110 words in base form. Of these, approximately 43 000 are non-compound base forms with paradigm information. The word list consists of words in citation form annotated with a paradigm and possibly a gradation pattern. There are 78 paradigms and 13 gradation patterns. For example, the entry for *käsi* 'hand' is *käsi 27* referring to paradigm 27 without gradation, whereas the word *pato* 'dam' is given as *pato 1F* indicating paradigm 1 with gradation

pattern F. From this description, a lexical transducer is compiled with a cascade of finite-state operations. For nominal paradigms, i.e. nouns and adjectives, inflection includes case inflection, possessive suffixes and clitics creating more than 2 000 word forms for each nominal. For the verbal inflection, all tenses, moods and personal forms are counted as inflections, as well as all infinitives and participles and their corresponding nominal forms creating more than 10 000 forms for each verb. In addition, the Finnish lexical transducer also covers nominal compounding.

This finite-state transducer lexicon was converted into an entry generator using the procedure outlined in Sect. 2.

4.2 Data Collections for Word Counts

To test the general applicability of our scoring methods, we decided to use two different data sources. The first data source is a large text data collection of Finnish and the second data source is the generally available search engine Google restricted to Finnish documents, which represents an even larger text collection.

The first data source is the *Finnish Text Collection*, which is an electronic document collection of the Finnish language. It consists of 180 million running text tokens. The corpus contains news texts from several current Finnish newspapers. It also contains extracts from a number of books containing prose text, including fiction, education and sciences. Gatherers are the Department of General Linguistics, University of Helsinki; The University of Joensuu; and CSCScientific Computing Ltd. The corpus is available through CSC [www.csc.fi]. We used this text collection to provide frequency counts of word forms.

The second data source, i.e. Google on Finnish documents currently[5] indexes approximately 152 million documents, which provided the document counts, i.e. they are not direct word frequency counts, but the word frequency is of course reflected in the number of documents that the word appears in.

4.3 Test Data Collections

To test how well the scoring methods are able to rank the best lexeme among the top lexeme candidates for a new and previously unseen word, we used two different test word collections, for which the correct base form and paradigm combinations had been determined manually.

To test the methods, we used the test data collection developed by Lindén [10], which is a set of word forms drawn from the *Finnish Text Collection*. In order to extract word forms that represent relatively infrequent and previously unseen words, 5000 word and base form pairs had been drawn at random from the frequency rank 100 001-300 000. To get new words, only word forms that were not recognized by the lexical transducer were kept. However, from this test data, strings containing numbers, punctuation characters, or only upper case

[5] February, 2009 by searching for *ja* 'and' in Finnish documents.

characters were also removed, as such strings require other forms of preprocessing in addition to some limited morphological analysis. Of the randomly selected strings, 1715 represented words not previously seen by the lexical transducer. For these strings, correct entries were created manually. Of these, only 48 strings had a verb form reading. The rest were noun or adjective readings. Only 43 had more than one possible reading.

A sample of the word forms from the first data set are: *ulkoasultaan* 'of its appearance', *kilpailulainsäädännön* 'legal framework on competition', *epätasa-arvoa* 'inequality', *euromaan* 'of a country using the euro', *työvoimapolitiikka* 'labor policy', *pariskunnasta* 'according to the married couple', *vastalausemyrskyn* 'of the objection storm', *liioitellun* 'of the exaggerated', *ruuanlaiton* 'of the cookery', *valtaannousun* 'of the ascent to power', *suurtapahtumaan*, *ostamiaan* 'the ones that they bought', ...

In Table 1, we see an example of the word form *ulkoasultaan* and the suggested base forms and paradigms as they have been generated by the entry generator and expanded with key word forms in order for a scoring method to determine the best lexeme for a morphological entry.

Table 1. Word form *ulkoasultaan* 'of its appearance' and the key word forms of the lexemes suggested by the entry generator. The English glosses are added in the table for readability.

ulkoasu	*1 noun*	'appearance'
		ulkoasu ulkoasun ulkoasua ulkoasuun ulkoasut
		ulkoasujen ulkoasuja ulkoasuihin
ulkoasu	*2 noun*	'appearance'
		ulkoasu ulkoasun ulkoasua ulkoasuun ulkoasut
		ulkoasujen∼ulkoasuitten∼ulkoasuiden
		ulkoasuja∼ulkoasuita ulkoasuihin
ulkoasullata	*73 I verb*	'to stuff from the outside'
		ulkoasullata ulkoasultaan ulkoasultasi ulkoasultaisi
		ulkoasullannee ulkoasullatkoon ulkoasullannut ulkoasullattiin
ulkoasu	*21 noun*	'appearance'
		ulkoasu ulkoasun ulkoasuta ulkoasuhun ulkoasut
		ulkoasuiden ulkoasuita ulkoasuihin

Using the entry generator developed by Lindén [10], we developed a larger second test data collection based on the words of the Finnish open source dictionary project Joukahainen [6]. We selected the words in the Joukahainen word list that were not included in the lexical data for the Finnish entry generator. Based on the existing lexical encoding of the Joukahainen project and the entry generator, the new words were encoded according to the guidelines of *The Research Institute for the Languages of Finland* [6] used in the entry generator. A

[6] http://joukahainen.puimula.org/

list of 11 026 new entries was created as test data. As the data came from an open-source word-list project, the words were all in base form.

4.4 Evaluation Measures, Baselines and Significance Test

We report our test results using recall and average precision at maximum recall. *Recall* means all the word forms in the test data for which an accurate base form suggestion is produced. *Average precision at maximum recall* is an indicator of the amount of noise that precedes the intended paradigm suggestions, where n incorrect suggestions before the m correct ones give a precision of $1/(n+m)$, i.e., no noise before a single intended base form per word form gives 100 % precision on average, and no correct suggestion at maximum recall gives 0 % precision. The *F-score* is the harmonic mean of the recall and the average precision. We will use only the recall and average precision among the top 6 candidates, as the output is intended for human post processing. In general, this will give us a lower, i.e. more conservative, recall than considering all candidates.

The random baseline for Finnish is that the correct entry is one out of 78 paradigms with one out of 13 gradations, i.e. a random correct guess would on the average end up as guess number 507.

As suggested by Lindén [10], we use the automatically derived entry generator from Sect. 4.1 as a baseline. Using his test data, the test results will be directly comparable to the baseline provided in Table 2 with recall 82 %, average precision 76 % and the F-score 79 %.

Table 2. Baseline for Finnish entry generator on infrequent word forms

Rank	Frequency	Percentage
#1	1140	66.5 %
#2	186	10.8 %
#3	64	3.7 %
#4	17	1.0 %
#5	4	0.2 %
#6	2	0.1 %
#7-∞	302	17.6 %
Total	1715	100.0 %

We also ran the entry generator directly on the base forms of our test data from the Joukahainen word collection in order to get the baseline provided in Table 3 indicating 66 % average precision and 72 % recall with an F-score of 69 %.

The significance of the difference between the baselines and the tested scoring methods is determined with matched pairs. The Wilcoxon Matched-Pairs Signed-Ranks Test indicates whether the changes in the ranking differences are statistically significant. For large numbers the test is almost as sensitive as the

Table 3. Baseline for Finnish entry generator on list of new base forms

Rank	Frequency	Percentage
#1	6043	54.8 %
#2	1196	10.8 %
#3	482	4.4 %
#4	157	1.4 %
#5	64	0.6 %
#6	11	0.1 %
#7-∞	3073	27.9 %
Total	11026	100.0 %

Matched-Pairs Student t-test even if it does not assume a normal distribution of the ranking differences.

5 Evaluation

We test how well the lexeme scoring methods outlined in Sect. 3 are able to select the best lexemes for a new word form using the test data described in Sect. 4.2. Word forms representing previously unseen words were used as test data in the experiment. The generated entries are intended for human post-processing, so the first correct entry suggestion should be among the top 6 candidates, otherwise the ranking is considered a failure.

If a ranking method ranks several candidates with the same score, a stable sorting algorithm will keep the original order of the lexemes. To test the effect of the proposed corpus-based ranking methods independently from the ordering given by the entry generator, the order of the entries generated from the entry generator were randomized before they were submitted to the corpus-method so as not to bias the corpus-method with the ranking order of the entry generator. We did five randomized runs for each evaluation and took the average of the ranking results.

In 5.1, we test the lexeme scoring with counts from the Finnish text corpus on the set of infrequent word forms. In 5.2, we test the scoring methods using page counts from the internet based on the Google search engine for the same set of word forms. In 5.3, we test the scoring methods on the set of new base forms using counts from the Finnish text corpus.

5.1 Corpus-based Lexeme Ranking of Word Forms

We evaluate the lexeme ranking method on base forms and paradigms generated by the lexicon-based entry generator from the test set of infrequent word forms using word counts from the *Finnish Text Collection* described in Sect. 4.2. In Table 4, we see the precision, recall and F-score of the three scoring methods.

Table 4. Precision, Recall and F-score of scoring methods

Method	Precision	Recall	F-Score
key word indicator	79 %	89 %	84 %
key word frequency	72 %	84 %	78 %
key word frequency with ilf	76 %	87 %	81 %

The *key word indicator scoring* performed best and ranked a correct entry among the top 6 candidates for 89 % of the test data as shown in Table 5, which corresponds to an average position of 1.96 for the first correct entry with 89 % recall and 79 % average precision, i.e. an 84 % F-score. All methods were statistically very highly significantly different using the corpus count data.

Table 5. Ranks of all the first correct lexeme suggestions using a text collection

Rank	Frequency	Percentage
#1	1184	69.0 %
#2	229	13.4 %
#3	71	4.1 %
#4	24	1.4 %
#5	15	0.9 %
#6	7	0.4 %
#7-∞	185	10.8 %
Total	1715	100.0 %

5.2 Page Count-based Lexeme Ranking of Word Forms

We also evaluate the lexeme ranking method on base forms and paradigms generated by the lexicon-based entry generator from the test set of infrequent word forms using the World-Wide Web page counts for pages retrieved over a period of some weeks from Google for key words of the paradigms. We retrieved the data from pages which Google gave a Finnish language code. We used this as a way to verify the ranking methods on an independent data collection. In Table 6, we see the precision, recall and F-score of the three scoring methods.

The *key word frequency scoring with inverse lexeme frequency* had the best overall performance and ranked a correct entry among the top 6 candidates for 83 % of the test data as shown in Table 7, which corresponds to an average position of 2.4 for the first correct entry with 83 % recall and 74 % average precision, i.e. an 78 % F-score. The difference to the pure frequency method was not statistically significant. However, the *key word indicator scoring* had the best recall of 88 %. The difference to the winning method was statistically significant on the lowest significance level.

Table 6. Precision, Recall and F-score of scoring methods

Method	Precision	Recall	F-Score
key word indicator	68 %	88 %	77 %
key word frequency	73 %	83 %	78 %
key word frequency with ilf	74 %	83 %	78 %

Table 7. Ranks of all the first correct lexemes using page counts from the World-Wide Web

Rank	Frequency	Percentage
#1	1114	65.0 %
#2	139	8.1 %
#3	63	3.7 %
#4	58	3.4 %
#5	32	1.9 %
#6	20	1.2 %
#7-∞	289	16.9 %
Total	1715	100.0 %

5.3 Corpus-based Lexeme Ranking of Base Forms

We evaluate the lexeme ranking method on base forms and paradigms generated by the lexicon-based entry generator for base forms in our subset of the Joukahainen word list using counts from the *Finnish Text Collection* described in Sect. 4.2. In Table 8, we see the precision, recall and F-score of the three scoring methods.

Table 8. Precision, Recall and F-score of scoring methods

Method	Precision	Recall	F-Score
key word indicator	56 %	78 %	65 %
key word frequency	50 %	71 %	59 %
key word frequency with ilf	51 %	72 %	59 %

The *key word indicator scoring* performed best and ranked a correct entry among the top 6 candidates for 78 % of the test data as shown in Table 9, which corresponds to an average position of 3.1 for the first correct entry with 78 % recall and 56 % average precision, i.e. an 65 % F-score. The indicator scoring was highly statistically significantly better than the two other scoring methods, which had no statistical difference between them.

Table 9. Ranks of all the first lexemes using counts from the text collection

Rank	Frequency	Percentage
#1	4198	38.0 %
#2	2038	18.5 %
#3	1018	9.2 %
#4	687	6.2 %
#5	381	3.5 %
#6	296	2.7 %
#7-∞	2418	21.9 %
Total	11026	100.0 %

5.4 Significance

All the ranking methods of the lexemes from the morphological entry generator were statistically highly significantly better than the random baselines according to the Wilcoxon Matched-Pairs Signed-Ranks Test. The difference between the two winning methods, i.e. the *key word indicator scoring* and the *key word frequency scoring with inverse lexeme frequency*, were statistically significant on all the test data collections.

The improvement in the recall of 6-7 % percentage points from the baseline model on two separate test data collections using the corpus word counts is also significant in practice as we counted words below the 6^th position as out-of-reach. This means that a significant number of additional correct classifications were now visible to the native speaker doing the entry revising.

6 Discussion

In this section, we discuss the results and give a brief overview of some related work. In 6.1, we compare our test results with previous efforts. In 6.2, we discuss some future work.

6.1 Discussion of Results

The problem when dealing with relatively low-frequency words is that an approach based on generating additional word forms of the lexemes may not contribute much. It may well be that the word form we are trying to classify is the only instance of the lexeme in the corpus. In that sense, turning to the internet for help seems like a good idea. It turned out that three times as many[7] of the generated word forms could be found on the internet as we were able to find in the 182 million word data collection of Finnish words.

It is interesting to note how the smaller amount of data affected the preference for the key word indicator scoring function. One reason for the good performance

[7] Data retrieved in December, 2008

of the word forms using the text collection is of course that the new words and the text collection were ideally matched, when the new words were drawn from the same collection as the counts. However, this is not a severe limitation. New words are usually drawn from known text collections. That is often one of the purposes for collecting the texts in the first place. However, also when the lexemes from an independent data collection are ranked with the same kind of word counts, it turns out that the key word indicator scoring is the most successful. As can be seen, the precision drops below that of the baseline suggesting that the additional lexemes that become visible to the lexicographer do not make it all the way to the top. Currently, we believe that this is due to the smaller size of the corpus and the fact that it may be slightly cleaner. In order for the frequency scoring to become more effective than the indicator scoring, we need a sufficiently large number of word forms for which a word count is available.

Using the Google data, the frequency ranking becomes more effective. The explanation for this seems to be that a larger number of word forms for each paradigm slightly reduces the distinctions between some of the best lexeme suggestions. This makes it interesting to use the more fine-grained frequency scoring which can be further scaled to emphasize the distinctiveness of the key word forms.

Sometimes a misspelling may have been more common than a correctly spelled word. E.g., the sixth highest scoring word in our material was *seuraavä*, with approx. 21 000 000 page counts, while its correctly spelled form, *seuraava*, had almost 500 000 page counts less. This was in most cases corrected by a higher average frequency of the remaining key word forms in the correct lexeme. Sometimes the incorrect lexemes happened to contain a homonym of some frequently occurring word, which raised the score of the lexeme above that of the correct lexeme candidate.

The fact that as a source for ranking entries, the corpus data fared slightly better than the internet may in our case also be attributable to the fact that Finnish word forms in the frequency range 100 000-300 000 may not be so rare after all due to the rich morphology and productive compounding mechanism of Finnish.

The larger test data collection was definitely less suited to the text collection. It also had a lower baseline to begin with indicating that base forms may not be the ideal words to classify with an entry generator that has been created for generating entries for any inflection. However, when constructing the test data we could benefit from the base forms by automatically discarding lexeme suggestions that did not have the input base form among the key word suggestions, e.g. if we know that we are looking for an entry for the English base form *swimming*, good lexeme candidates like *swim V* can be mechanically discarded in favor of *swimming N*. This option was not used when testing the performance of the ranking methods, as we wanted to evaluate their contributions independently.

It remains to be seen how the base form data collection would perform using Google page counts. Would the higher frequency counts for more key word forms favor a more sensitive scoring method in the same way as it did on the smaller

test data collection? As the methods were statistically significantly different on the smaller data set, we have reason to believe that this prediction will hold.

From a practical point of view, we were able to significantly reduce the workload of encoding lexical entries as much of the task can now be accomplished automatically by suggesting only those lexemes for which there is at least some support in the corpus in addition to the word form being investigated. However, a significant practical change is that assigning base forms and paradigms to words which previously required an expert lexicographer can now be accomplished by a native speaker making a choice between a very limited set of lexeme suggestions.[8]

6.2 Comparison with similar or related efforts

A related idea of expanding key word forms of paradigms to identify new words and their paradigms has been suggested by Hammarström & al [2]. However, their approach was to automatically deduce rules for which they could find as much support as was logically possible in order to make a safe inference. This leads to safely extracting words that already have a number of word forms in the corpus, i.e. mid or high-frequency words, which for all practical purposes have most likely already been encoded and are readily available in public domain morphological descriptions like the Ispell dictionaries [7] or more advanced descriptions like the Finnish dictionary *Kotimaisten kielten tutkimuskeskuksen nykysuomen sanalista* [6]. It should be noted that Hammarström & al [2] drew the conclusion that it is recommendable that a linguist writes the extraction rules. In addition, they used an even smaller corpus than we did. This indicates that their reliance on an approach that is most similar to our *key word indicator scoring* will probably be less effective when the corpus-size increases, requiring their linguist to write additional constraints for the extraction rules.

The approach suggested by Mikheev[13, 14] aims at solving the issue of unknown words in the context of part-of-speech taggers. However, in that context the problem is slightly simpler as the guesser only needs to identify a likely part of speech, not the base form and the full inflectional paradigm of a word. He suggests an automatic way of extracting prefix and postfix patterns for guessing the part of speech. A related approach aiming at inducing paradigms for words and inflectional morphologies for 30 different languages is suggested by [18].

Since there is a growing body of translated text even for less studied languages, there are interesting approaches using multi-lingual evidence for inducing morphologies, see e.g. Yarowsky and Wicentowski[19]. This approach is particularly fruitful if we can benefit from the similarities of closely related languages.

If we cannot find enough support for any particular lexeme for a given word form, e.g. if the word is too infrequent so that there are no other inflections than the original word form, we need a way to make inferences based on related or similar strings. We need to make inferences based on the analogy with already known words as suggested e.g. by Goldsmith [3] or Lindén [9–11].

[8] For a demo of the classification interface, see http://www.ling.helsinki.fi/cgi-bin/omor/omorfi-cgidemo.py

6.3 Future Work

The current study aimed at evaluating the effect of corpus evidence in isolation. We have reason to believe that an approach which combines the output of the entry generator with the methods evaluated in this article, e.g. by relying on the ordering suggested by the entry generator when the corpus evidence does not distinguish between the ranks of the lexeme suggestions is an effective way of combining the corpus evidence with the entry generator. This essentially means that the entry generator functions as fall-back when there is a lack of corpus evidence.

Currently we only extract inflectional information in the form of lexical entries, even if the context of a new word also contributes other types of lexical information such as part of speech, argument structure and other more advanced types of syntactic and semantic information.

It is important to note that our experiment verifies that the lexeme ranking can be performed using page counts instead of word counts with a sufficiently large document corpus, which is by no means self-evident. Many of the word forms will refer to the same pages, which also opens up avenues for future research. One could perhaps use page counts for a combination of the base form with some other word form of a lexeme in order to reduce the noise by searching directly for pages with combinations of several key word forms.

The Internet in addition to page counts also provides some context for a word form. Essentially this means that we have access to the local semantic context of a word. The Internet is an ever-changing medium and any linguistic data derived form it is subject to change. The challenge is to harness this evidence to distill the information inherent in large numbers while still adapting to the significant changes in language use.

7 Conclusions

We have proposed and successfully tested new methods for ranking lexemes generated for word forms of new words using additional corpus evidence for key word forms of the lexemes suggested by an entry generator. We tested the model on Finnish, which is a highly inflecting language with a considerable set of inflectional paradigms and stem change categories. A key finding was that the ranking functions that can better take into account fine-grained words counts seems likely to perform better on larger and perhaps inevitably more noisy corpora. We tested the effects of the methods independently from the prior ranking by the entry generator. The methods can be combined to provide optimal performance.

Our corpus-based ranking methods have 56-79 % average precision and 78-89 % recall among the top 6 candidates, i.e. an F-score of 65-84 %, indicating that the first correct entry suggestion is on the average found in positions 1.9-3.1. The ranking methods based on corpus evidence were found to be significant in practice by increasing the recall among the top 6 candidates with 7-8 %, which

saves the lexicographer some work when reducing the need to create entries from scratch.

Acknowledgments

We are grateful to the Finnish Academy. We are also grateful to Inari Listenmaa, who used her native language skills to evaluate the entry generator by revising entries for new words from the Joukahainen word list.

References

1. Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saralar, M., Stolcke, A. 2007 Morph-based speech recognition and modeling of out-of vocabulary words across languages. In ACM Transactions on Speech and Language Processing, 5(1) article 3.
2. Forsberg, M., Hammarström, H., and Ranta, A. 2006. Morphological Lexicon Extraction from Raw Text Data. FinTAL 2006, LNCS 4139, pages 488499, 2006.
3. Goldsmith, J. A. 2007. Morphological Analogy: Only a Beginning, http://hum.uchicago.edu/ jagoldsm/Papers/analogy.pdf
4. HFST–Helsinki Finite-State Technology. 2008. http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/index.shtml
5. Koskenniemi, K.. 1983. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. PhD Thesis. Department of General Linguistics, University of Helsinki, Publication No. 11.
6. Kotimaisten kielten tutkimuskeskuksen nykysuomen sanalista, 2007. Research Institute for the Languages of Finland. http://kaino.kotus.fi/sanat/nykysuomi/
7. Kuenning, G. 2007 Dictionaries for International Ispell, http://www.lasr.cs.ucla.edu/geoff/ispelldictionaries.html
8. Kurimo, M., Creutz, M., Turunen, V. 2007. Overview of Morpho Challenge in CLEF 2007. In Working Notes of the CLEF 2007 Workshop, pages. 19-21.
9. Lindén, K. 2008. A Probabilistic Model for Guessing Base Forms of New Words by Analogy. In 9th International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel, LNCS, vol. 4919, pages. 106-116. Springer.
10. Lindén, K. 2009. Guessers for Finite-State Transducer Lexicons. In CICling-2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics, March 1-7, 2009, Mexico City, Mexico.
11. Lindén, K. 2009. Entry Generation by Analogy–Encoding New Words for Morphological Lexicons. In Northern European Journal of Language Technology. May, 2009.
12. Lingsoft, Inc.: Demos, http://www.lingsoft.fi/?doc_id=107&lang=en
13. Mikheev, A. 1996. Unsupervised Learning of Word Category Guessing Rules. In: Proc. of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96), pages 327-334.
14. Mikheev, A. 1997. Automatic Rule Induction for Unknown-Word Guessing. In Computational Linguistics,. 23(3), pages 405-423.
15. Oflazer, K., Nirenburg, S., McShane, M. 2001. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. In Computational Linguistics, 27(1), pages 59-85.

16. Pirinen, T. 2008. Open Source Morphology for Finnish using Finite-State Methods (in Finnish). Technical Report. Department of Linguistics, University of Helsinki.
17. Sakarovitch, J. 2003. éléments de théorie des automates.
18. Vuibert Wicentowski, R. 2002. Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework. PhD Thesis, Baltimore, USA.
19. Yarowsky, D. and Wicentowski. R. 2000. Minimally Supervised Morphological Analysis by Multimodal Alignment. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.