Northern European Journal of Language Technology, 2009, Vol. 1, Article 1, pp 1–25
DOI 10.3384/nejlt.2000-1533.09111

Entry Generation by Analogy – Encoding New Words for Morphological Lexicons

Krister Lindén

University of Helsinki
Department of General Linguistics
Krister.Linden@helsinki.fi

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 lexicon, we need to indicate their base form and inflectional paradigm. In this article, we evaluate a combination of corpus-based and lexicon-based methods for assigning the base form and inflectional paradigm to new words in Finnish, Swedish and English finite-state transducer lexicons. The methods have been implemented with the open-source *Helsinki Finite-State Technology (Lindén & al., 2009)*. 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 combining the probabilities calculated from corpus data and from lexical data, we get a more precise combined model. The combined method has 77-81 % precision and 89-97 % recall, i.e. the first correctly generated entry is on the average found as the first or second candidate for the test languages. A further study demonstrated that a native speaker could revise suggestions from the entry generator at a speed of 300-400 entries per hour.

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 (1996, 1997) noted 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. (2001) 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 to which paradigm a given citation form belongs.

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 (2002) and Goldsmith (2007). 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 (Creutz & al., 2007). For a comparison of some recent successful segmentation methods, see the Morpho Challenge (Kurimo & al., 2007).

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 analyzes in the context of more frequent words. Especially for Germanic and Finno-Ugric languages, there are already large-vocabulary descriptions available and new words tend to be compounds of acronyms and loan words with existing words. 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 the compounds. It has previously been demonstraed by Lindén (2008a) that already training sets as small as 5000 inflected word forms and their manually determined base forms will give a reasonable result for guessing base forms of new words by analogy, which was tested on a set of languages from different language families, i.e. English, Finnish, Swahili and Swedish.

In addition, there are a host of large but shallow hand-made morphological descriptions available, e.g., the Ispell collection of dictionaries (Kuenning, 2007) for spell-checking purposes, and many well-documented morphological analyzers are commercially available, e.g. Lingsoft (2008). It has also been demonstrated by Lindén (2009) 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.

In this work, we propose and evaluate a new method for analogically *determining the base form and inflectional paradigm of inflected forms* of new words by using both corpus-based and lexicon-based information. In Section 2, we outline the directly related previous work. In Section 3, we describe the new method. In Section 4, we present the training and test data. In Section 5, we evaluate the model. In Section 6, we discuss the method and the test results in light of the existing literature on analogy.

2 Previous Work

2.1 Corpus-based Base Form Guesser

Assume that we have a set of words, $w \in W$, from a text corpus for which we have determined the base forms, $b(w) \in B \subset W$, i.e. the lexicon look-up form[1]. In addition, we have another set of words, $o \notin W$, for which we would like to determine the most likely base forms, $b(o)$. For this purpose, we use the analogy that w is to o as $b(w)$ is to $b(o)$. This relationship is illustrated in Figure 1.

w	:	o		*kokeella*	:	*aikeella*
↕		↕		↕		↕
$b(w)$	:	$b(o)$		*koe*	:	?

Fig. 1. The analogy w is to o as $b(w)$ is to $b(o)$ illustrated by the Finnish words *kokeella* ‘with the test’ and *aikeella* ‘with the intention’.

We use the analogical relation for deriving transformations $w \rightarrow b(w)$ from the differences between the known word and base forms. The transformations can then be applied to a new word o in order to generate a base form that should be similar to an existing base form $b(w)$. Several transformations may apply to any particular o and we wish to determine the most

[1] The lexicon look-up form may in some languages be a root that is not used in running text. However, for the purpose of analogy, it is sufficient that the lexicon look-up form has been determined and added to the set of words, $w \in W$.

likely $b(o)$ in light of the evidence, i.e. we wish to find the $b(o)$, which maximizes the joint probability $P(o, w \rightarrow b(w), b(w), b(o))$ for the new word o.

The joint probability cannot be directly estimated from a corpus. However, we may assume that $b(w)$ and o are independent of one another and that the probability of o is constant during the maximization. As an additional simplification, we can assume no knowledge of the distribution[2] of the analog base forms, $b(w)$. By applying the chain rule to the joint probability and using the simplifying assumptions and the standard Viterbi approximation, Lindén (2008a) shows how to arrive at Equation 1:

$$\underset{b(o)}{\operatorname{argmax}} P(w \rightarrow b(w) \mid o) P(b(o) \mid b(w)) \quad (1)$$

The equation can be instantiated with a model for estimating the parameters from a character-aligned corpus of word and base form pairs creating a probabilistic base form guesser. For aligning words pairs, see e.g. (Lindén, 2006).

For the experiment the guesser was instantiated with a symmetrical model for concatenating morphology which is sufficient for a large set of languages. The model was successfully tested on Finnish, Swedish, English and Swahili (Linden, 2008a). For an illustration of the model, see Figure 2.

$$\begin{array}{ccc} \alpha \mathrm{X} \gamma & : & \alpha \Psi \gamma \\ \updownarrow & & \updownarrow \\ \beta \mathrm{X} \delta & : & \beta \Psi \delta \end{array},$$

Fig. 2. An instance of the analogy model, where stems X and Ψ are any strings, and $\alpha \rightarrow \beta$ is a prefix transformation and $\gamma \rightarrow \delta$ a suffix transformation.

The likelihood of the stems X and Ψ are proportional to their lengths[3], whereas the likelihoods of the prefix $\alpha \rightarrow \beta$ and suffix transformation $\gamma \rightarrow \delta$ are estimated from a corpus[4].

2.2 Lexicon-based Entry Generator

Assume that we have a finite-state transducer lexicon T which relates base forms, $b(w)$, to inflected words, w. Let w belong to the input language L_I and $b(w)$ to the output language L_O of the transducer lexicon. Our goal is to create an entry generator for inflected words that are unknown to the lexicon, i.e. we wish to provide the most likely base forms $b(u)$, e.g. *b(u)='warble',* for an unknown input word *u='warbled',* $u \notin L_I$. 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. Sakarovitch (2003).

We can regard the left quotient as the set of postfixes of language L_1 that complete words from a language L_2, such that the resulting word is in language L_1. If L_1 and L_2 are formal

[2] One could argue that the a priori probability of a base form correlates with how likely a concept is as a basis for an analogy, e.g. rare base forms by definition rarely reinforce our perception of suitable base forms.

[3] The motivation for having the word length proportional to the word probability is that generating a new stem essentially presupposes an almost random choice from an alphabet for each additional character. This is also consistent with Zipf's second law that the negative log probability of the word is proportional to the word length as pointed out by Yves Lepage.

[4] From an analyzed and aligned corpus of word forms and base forms, it is possible to extract e.g. suffix and prefix transformations. For each aligned word form substring, it is possible to estimate the probability distribution of the base form substrings it corresponds to.

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:

$$L_1 \setminus L_2 = \{a \mid \exists x((x \in L_2) \wedge (xa \in L_1))\} \quad (2)$$

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 weight $p(w)$ 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:

$$W = \{w \mid \exists w(w \in \Sigma)\} \text{ with weights } p(w) = C \mid w \mid, \quad (3)$$

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 for the universal language U of the pair alphabet of T:

$$G(T) = W\ T \setminus \mathrm{U} \quad (4)$$

Theorem. Let T be a weighted lexical transducer and W a weighted universal language with transition weights $\omega + \delta$, where $\delta > 0$ and ω the maximum of any transition weight in T. Then $G(T) = W\ T \setminus \mathrm{U}$ is the longest matching suffix entry generator (Lindén, 2009).

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

3. Methodology

Assume that we have a probabilistic utility for guessing base forms from inflected forms, *ProbabilisticBaseFormGuesser* (Lindén, 2008a). Also assume that we are able to create a set of base forms which are classified according to their inflectional paradigms, from which we derive a paradigm classifier for base forms, *BaseFormParadigmGuesser* (Lindén, 2008b, 2009). For a cascade of analogies used in a base form and paradigm classifier, see Figure 3.

w	:	o	*kokeella*	:	*aikeella*
↕		↕	↕		↕
$b(w)$	:	$b(o)$	*koe*	:	*aie*
$b(w)$	:	$b(o)$	*koe*	:	*aie*
↕		↕	↕		↕
$e(b(w))$	:	$e(b(o))$	*koe* 48*D*	:	?

Fig. 3. The cascade of analogies where w is to o as $b(w)$ is to $b(o)$, and $b(w)$ is to $b(o)$ as $e(b(w))$ is to $e(b(o))$ as illustrated by the Finnish words *kokeella*[5] 'with the test' and *aikeella* 'with the intention'.

[5] *koe 48 D* means that *koe* 'test' belongs to the 48th paradigm and has stem gradation pattern D, i.e.'k' ↔ ' '. Stem gradation indicates that the stem has different lengths in the stem consonant pattern according to the gradation triggered by the inflected form. A long pattern is called strong gradation and a short is called weak.

The paradigm classifier generates possible paradigm candidates for a base form by analogy. The two components can be used for creating a probabilistic entry generator by taking an inflected word form, *WordForm*, transforming it to a set of base forms with *ProbabilisticBaseFormGuesser*, and guessing their inflectional paradigms using the *BaseFormParadigmGuesser*. Entries generated from a word form by a cascade of probabilistic models encoded as weighted transducers are characterized by Equation 5.

$$CorpusModel = WordForm \circ ProbabilisticBaseFormGuesser \circ BaseFromParadigmGuesser \quad (5)$$

As a baseline, we use the entry generators that have been automatically derived from full-scale transducer lexicons, *LexicalEntryGenerator* (Lindén, 2009). The entry generators take an inflected word form, *WordForm*, and produce a set of base forms with inflectional paradigms. Entries generated from a word form by a lexicon-based model encoded as a weighted lexical transducer are characterized by Equation 6.

$$LexicalModel = WordForm \circ LexicalEntryGenerator \quad (6)$$

We combine the corpus-based model and the lexicon-based model by taking the surface projection of the output side of the weighted transducers and intersecting them in Equation 7.

$$CombinedModel = proj(CorpusModel) \,\&\, proj(LexicalModel) \quad (7)$$

In 3.1, we outline the implementation of the corpus-based paradigm guesser, and in 3.2, we describe the combination of the baseline model with the corpus-based paradigm guesser. All models have been implemented using our open-source *Helsinki Finite-State Technology* (Lindén & al. 2009) which can be freely downloaded from our web site (HFST, 2008).

3.1 Corpus-based Entry Generator

As the model for analogy between word forms and base forms is general, we can repeat the analogy when going from base forms to entries. We take the entries $e(b(w))$ for base forms $b(w)$ in a lexicon as our starting point and derive a suffix model for the analogy between base forms and lexical entries. See illustration in Figure 4.

$$\begin{array}{ccc} b(w) & : & b(o) \\ \updownarrow & & \updownarrow \\ e(b(w)) & : & e(b(o)) \end{array} \qquad\qquad \begin{array}{ccc} koe & : & aie \\ \updownarrow & & \updownarrow \\ koe\,48D & : & ? \end{array}$$

Fig. 4. The analogy $b(w)$ is to $b(o)$ as $e(b(w))$ is to $e(b(o))$ as illustrated by the Finnish words *koe* ‘test’ and *aie* ‘intention’.

From Equation 1, we derive Equation 8,

$$\underset{e(b(o))}{\operatorname{argmax}} P\big(b(w) \rightarrow e(b(w)) \mid b(o)\big) P\big(e(b(o)) \mid e(b(w))\big) \quad (8)$$

which can be instantiated in the same way as the original model with the additional simplification that paradigm information is added only to the end of the entry resulting in the *BaseFormParadigmGuesser*.

Both models are probabilistic and there is no particular need to extract the base forms as intermediate results, so the models can be combined sequentially into one single model implemented as a cascade of weighted finite-state transducers. The combined model yields the entry $e(b(o))$ for an inflected form o according to Equation 9,

$$\underset{e(b(o))}{\operatorname{argmax}} \begin{pmatrix} P(w \rightarrow b(w) \mid o)\, P(b(o) \mid b(w)) \times \\ P(b(w) \rightarrow e(b(w)) \mid b(o))\, P(e(b(o)) \mid e(b(w))) \end{pmatrix} \tag{9}$$

We instantiate the *BaseFormParadigmGuesser* with a suffixing model for adding paradigm information to base forms. For an illustration of the model, see Figure 5.

The likelihoods of the stems X and Ψ are proportional to their lengths, whereas the likelihoods of the suffix transformations $\gamma \rightarrow \delta$ is estimated from a corpus. The two analogy models are cascaded in order to create base forms with paradigm information from inflected word forms.

$$\begin{array}{ccc} \mathrm{X}\gamma & : & \Psi\gamma \\ \updownarrow & & \updownarrow \\ \mathrm{X}\delta & : & \Psi\delta \end{array},$$

Fig. 5. An instance of the analogy model, where stems X and Ψ are any strings, and $\gamma \rightarrow \delta$ a suffix transformation.

3.2 Combined Entry Generator

The corpus-based model needs more training data the more complex it becomes, i.e. the more linguistic parameters it has that need to be estimated from the corpus. In order to keep the set of training data within reasonable limits, we do not wish to add too much linguistic complexity into the corpus-based model. Instead we rely on the lexical-model for encoding prior linguistic knowledge.

In Section 3.1, we explicitly extract the entries from the corpus-based model. However, we can store the entries in closed form in a weighted automaton. In Section 2.2, we described a lexicon-based entry generation model, i.e. the baseline method, which can be derived from a finite-state transducer lexicon of any language. The results from this model can also be stored in closed form in a weighted automaton. By combining the two results using automata intersection or transducer composition, we get a set of entries in closed form combining the likelihood estimates of both the corpus-data and the lexicon-data. The entry candidates can be extracted with an n-best algorithm.

Generally, one can characterize the corpus-based entry generator, *CorpusModel*, as inducing a likelihood ordering over the possible base-forms for an inflected form, whereas the lexicon-based entry generator, *LexicalModel*, promotes entries that are paradigmatically motivated by the lexicon. By combining the probabilities calculated from the corpus data and the lexicon data, we get a more precise *CombinedModel* with the added benefit that we have one repository for statistical data and another for linguistic knowledge, which can both be updated independently.

4 Data Sets

To test the entry generator for finite-state transducer lexicons, we created transducer lexicons from existing lexical resources for three different languages: English, Finnish and Swedish using the *Helsinki Finite-State Technology* (HFST, 2008; Lindén & al., 2009). In 4.1, we describe the lexical resources and outline the procedure for creating the finite-state transducer lexicon. Words unknown to the lexicons were drawn from three language-specific text collections. We manually determined the correct entries for a sample of the unknown words. In 4.2, we describe the text collections and the sample used as test data. In 4.3, we describe the evaluation method and characterize the baselines.

4.1 Lexical Data for Finite-State Transducer Lexicons

Lexical descriptions relate look-up words, e.g. base forms, to other words, e.g. inflected word forms, and indicate their relation to the look-up word. In a morphological lexicon, this relation can be described either as a base form classified with a paradigm and a set of derivation rules for the word forms of the paradigm, or it can be described as a full-form lexical description with a list of all the inflected forms of each base form. Regardless of the initial form of the lexical description, the final morphological lexicon can be implemented with finite-state transducer technology. The morphological finite-state transducer lexicon relates a word in dictionary form to all of its inflected forms. For an introduction, see e.g. Koskenniemi (1983). Essentially this means that composing the transducer lexicon with one inflected word form will extract a new transducer containing the possible base forms with morphological tags to indicate how the inflected word form is related to the base form.

A weighted finite-state transducer lexicon can contain weights for many purposes. From our perspective, a useful set of weights are estimates of the relative frequency of the word forms encoding their a priori likelihoods. Acquiring such estimates requires a disambiguated corpus. As we only have lexical descriptions and, assuming that there are or we have created inflectional paradigms for each word in the lexicon, we can estimate the relative frequency of the paradigms. It has also been demonstrated by Karlsson (1992) that it is preferable to have as few parts as possible in a multipart morphological compound analysis. For lack of better a priori estimates, the weighted finite-state lexicon-based transducer lists the morphological analyses primarily according to the number of analyzed compound parts and secondarily in paradigm frequency order.

Most languages have ready-made inflectional paradigms as their lexical description. From this a finite-state transducer lexicon can be manually compiled. However, for languages which typically have few inflected forms for each base form, it is feasible to have a full-form description of all the lexical entries. If we only have a full-form lexical description as a starting point, we still need to induce paradigms in order to be able to generate lexical entries with complete sets of forms for new words.

English. For English we use *FreeLing 2.1* (2007). The FreeLing English lexical resource was automatically extracted from WSJ, with some manual post-editing and completion of the lexical entries. It contains about 55 000 word forms corresponding to some 40 000 different combinations of base form and part-of-speech. For each part-of-speech, English has only a small set of forms, which may be further restricted due to phonological or semantic reasons. In this particular case, the set of forms may also be restricted by the fact that the form did not occur in the Brown corpus.

We induce paradigms from the full-form lexical description for English in the following manner: we automatically align the characters of the base form and the inflected forms and

determine the longest common prefix for the base form and all the inflected forms. The remaining set of endings with morphological tags, possibly with some characters from the stem, is considered a paradigm. Since some words may have individual patterns with missing forms, the automatically induced set of paradigms becomes relatively large. We get 489 paradigms for English out of which only 151 occur more than once in a Zipf-like distribution.

Finnish. In order to create the Finnish dictionary, we used the Finnish word list *Nykysuomen sanalista* (2007), which contains 94 110 words in base form. Of these, approximately 43 000 are non-compound base forms classified with paradigm information. The word list consists of words in citation form annotated with paradigm and 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 (Pirinen, 2008). For nominal paradigms, 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.

Swedish. For Swedish we use the open source full-form dictionary *Den stora svenska ordlistan* (Westerberg, 2008) with approximately 55 000 words in base form. For each base form, the part of speech is given. For each part-of-speech, there is a given set of inflected forms, e.g. for nouns there are always eight forms, i.e. all combinations of singular and plural, nominative and genitive, definite and indefinite forms. For any word form, there may be an empty slot, if the form is considered non-existent for some reason, e.g. phonologically or semantically. In addition, each word may have an indication of whether it can take part in compounding which is prolific in Swedish.

We use the same procedure for inducing paradigms for Swedish as we used for English. We get 1333 paradigms out of which 544 occur more than once in a Zipf-like distribution.

4.2 Test Data

A set of previously unseen words in inflected form serves as the test words, for which we wish to determine their base form and inflectional paradigm. In order to extract word forms that represent relatively infrequent and previously unseen words, we used various text collections for English, Finnish and Swedish. We draw 5000 words at random from the frequency rank 100 001-300 000 as test material for each language. The most frequent word form has rank 1. The max rank for a language is the rank of its most infrequent inflected word form. Since we are interested in new words, we only count inflected forms that are not recognized by the lexical transducers we have created. However, from the test data, we remove strings with numbers, punctuation characters, or only upper case characters, as such strings also require other forms of preprocessing in addition to some limited morphological analysis.

English. For English, we used part of *The Project Gutenberg* text collection, which consists of thousands of books. For this experiment we used the English texts released in the year 2000 [http://www.gutenberg.org/]. The tokens consist of 266 000 inflected forms of 175 000 base forms.

Of the selected strings, 3100 represented words not previously seen by the lexical transducer. For these strings, correct entries were created manually for the first 25 %, i.e. 775

new entries. Of these, 60 strings had verb form readings, 610 noun readings and 161 adjective readings, and 14 adverb readings. Only 79 strings had more than one reading.

A sample of test strings are: *florin, disfranchised, chimney-pieces, Beechwood, warbled, sureness, sitting-rooms, marmoset, landscape-painter, half-burnt, Burlington, …*

Finnish. For Finnish, we used 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 CSC–Scientific Computing Ltd. The corpus is available through CSC [www.csc.fi]. The tokens consist of 4 million inflected forms from 1.8 million base forms.

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 test strings are: *ulkoasultaan* 'by the appearance'*, kilpailulainsäädännön* 'of the competition legislation'*, epätasa-arvoa* 'unequality'*, työvoimapolitiikka* 'labour policy'*, pariskunnasta* 'from the married couple'*, vastalausemyrskyn* 'of the objection storm'*, ruuanlaiton* 'of the cooking'*, valtaannousun* 'of the ascent to power'*, suurtapahtumaan* 'for the mega-event'*, ostamiaan* 'the ones that they had bought*, …*

Swedish. For Swedish, we used the *Finnish-Swedish Text Collection*, which is an electronic document collection of the Swedish language of the Swedish speaking minority in Finland. It consisted of 35 million tokens. The corpus contains news texts from several current Finnish-Swedish newspapers. It also contains extracts from a number of books containing fiction prose text. Gatherers are The Department of General Linguistics, University of Helsinki; CSC–Scientific Computing Ltd. The corpus is available through CSC [www.csc.fi]. The tokens consist of 765 000 inflected forms of 445 000 base forms.

Of the selected strings, 1756 represented words not previously seen by the lexical transducer. For these strings, correct entries were created manually for first 25 %, i.e. 439 new entries. Of these, 37 strings had a verb form reading, 387 noun readings, 47 adjective readings. Only 48 strings had more than one reading.

A sample of the test strings are: *finrummet* 'the salon'*, chansons* 'chansons'*, översvämmande* 'inundating'*, tonsiller* 'tonsils'*, sjöfartspolitiska* 'of the maritime policy'*, reliken* 'the relic'*, oskött* 'unattended'*, antidopingkommitté* 'anti-doping committee'*, …*

4.3 Evaluation Measures, Baselines and Significance Test

We report our test results using recall and average precision at maximum recall. *Recall* means all the inflected 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 base form suggestions. For *n* incorrect suggestions before the *m* correct ones, we get a precision of *1/(n+m)*. If we have no noise before a single intended base form per word form, we get 100 % precision on average, and if we have no correct suggestion at maximum recall, we get 0 % precision. As only a small percentage of the test data have more than one possible outcome, we will use the first correct result for counting the average precision, i.e. *1/(n+1)*. The *F-score* is the harmonic mean of the recall and the average precision.

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. For English, an average random guess ends up in position 245 and, for Swedish, in position 667.

As suggested by Lindén (2009), we use the automatically derived entry generators in Section 2.2 as baselines. Using his test data, the test results will be directly comparable to the baselines provided in Table 1.

Table 1. Baselines for Finnish, English and Swedish entry generators.

Language	*Recall*	*Precision*	*F-score*
English	0.83	0.72	0.78
Finnish	0.82	0.76	0.79
Swedish	0.87	0.71	0.78

The significance of the difference between the baselines and the proposed methods is tested 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 Matched-Pairs Student t-test even if it does not assume a normal distribution of the ranking differences.

5 Experiments

We test how well the entry generators outlined in Section 3 are able to predict the correct base form and paradigm for an inflected word form using the test data described in Section 4. Of the randomly chosen strings from the test data range, 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 few candidates, otherwise the ranking is considered a failure. We chose to study the top six candidates[6]. Homonyms and loan words may sometimes have several inflectional paradigms and therefore more than one suggestion may be correct. However, in the experiments only the first correct suggestion is counted. In 5.1, we test the corpus-based entry generators separately, and in 5.2, we test them in combination with the lexicon-based entry generators. In 5.3, we evaluate the significance of the test results.

5.1 Corpus-based Entry Generators

We evaluate the corpus-based entry generators separately for each test language.

English. The English entry generator generated a correct entry among the top 6 candidates for 80 % of the test data as shown in Table 2 with an average position of 3.1 for the first correct entry with 80 % recall and 56 % average precision, i.e. a 66 % F-score.

Finnish. The Finnish entry generator generated a correct entry among the top 6 candidates for 80 % of the test data as shown in Table 3 with an average position of 2.6 for the first correct entry with 80 % recall and 69 % average precision, i.e. a 74 % F-score.

[6] In general, 5-7 entities at a glance is the maximal cognitive load a human user is comfortable with. In addition, entries suggested below the 6th position rarely contribute correct results.

Table 2. Ranks of all the first correct entries by the English entry generator.

Rank	*Freq*	*Percentage*
#1	282	36.4 %
#2	147	19.0 %
#3	94	12.1 %
#4	54	7.0 %
#5	26	3.4 %
#6	16	2.1 %
#7-∞	156	20.1 %
Total	775	100.0 %

Table 3. Ranks of all the first correct entries by the Finnish entry generator.

Rank	*Freq*	*Percentage*
#1	984	57.4 %
#2	193	11.3 %
#3	95	5.5 %
#4	56	3.3 %
#5	35	2.0 %
#6	15	0.9 %
#7-∞	337	19.7 %
Total	1715	100.0 %

Swedish. The Swedish entry generator generated a correct entry among the top 6 candidates for 86 % of the test data as shown in Table 4 with an average position of 2.3 for the first correct entry with 86 % recall and 72 % average precision, i.e. a 78 % F-score.

Table 4. Ranks of all the first correct entries by the Swedish entry generator.

Rank	*Freq*	*Percentage*
#1	259	59.0 %
#2	66	15.0 %
#3	26	5.9 %
#4	15	3.4 %
#5	9	2.1 %
#6	3	0.7 %
#7-∞	61	13.9 %
Total	439	100.0 %

5.2 Combined Entry Generators

We evaluate each corpus-based entry generator in combination with its baseline model, i.e. combined with its lexicon-based entry generator.

English. The combined English entry generator generated a correct entry among the top 6 candidates for 97 % of the test data as shown in Table 5 with an average position of 1.56 for the first correct entry with 97 % recall and 81 % average precision, i.e. a 88 % F-score. Note that for English, all results below the 4th position are considered inaccessible and are counted as rank #7 or lower because the impression from a small test sample was that they were mostly incorrect and more confusing than helpful to a human reader.

Table 5. Ranks of all the first correct entries by the combined English entry generator.

Rank	*Freq*	*Percentage*
#1	522	67.4 %
#2	175	22.6 %
#3	52	6.7 %
#4	0	0.0 %
#5	0	0.0 %
#6	0	0.0 %
#7-∞	26	3.4 %
Total	775	100.0 %

Finnish. The combined Finnish entry generator generated a correct entry among the top 6 candidates for 89 % of the test data as shown in Table 6 with an average position of 1.95 for the first correct entry with 89 % recall and 81 % average precision, i.e. a 85 % F-score.

Table 6. Ranks of all the first correct entries by the combined Finnish entry generator.

Rank	*Freq*	*Percentage*
#1	1264	73.7 %
#2	131	7.6 %
#3	50	2.9 %
#4	47	2.7 %
#5	20	1.2 %
#6	11	0.6 %
#7-∞	186	10.8 %
Total	1715	100.0 %

Swedish. The combined Swedish entry generator generated a correct entry among the top 6 candidates for 93 % of the test data as shown in Table 7 with an average position of 2.03 for the first correct entry with 93 % recall and 77 % average precision, i.e. a 84 % F-score.

Table 7. Ranks of all the first correct entries by the combined Swedish entry generator.

Rank	*Freq*	*Percentage*
#1	289	65.8 %
#2	53	12.1 %
#3	27	6.2 %
#4	11	2.5 %
#5	14	3.2 %
#6	14	3.2 %
#7-∞	31	7.1 %
Total	439	100.0 %

Table 8. Baselines and results for the corpus-based and combined entry generators for English, Finnish and Swedish.

	Lexical Baseline			*Corpus-based Model*			*Combined Model*		
Language	*Recall*	*Precision*	*F-score*	*Recall*	*Precision*	*F-score*	*Recall*	*Precision*	*F-score*
English	0.83	0.72	0.78	0.80	0.56	0.66	0.97	0.81	0.88
Finnish	0.82	0.76	0.79	0.80	0.69	0.74	0.89	0.81	0.85
Swedish	0.87	0.71	0.78	0.86	0.72	0.78	0.93	0.77	0.84

5.3 Significance

All the combined morphological entry generators were statistically highly significantly better than their lexical baseline according to the Wilcoxon Matched-Pairs Signed-Ranks Test. Despite the fact that the corpus-based entry generators score below their respective baselines, as can be seen in Table 8, they contributed an important source of probabilistic information when ranking the entry candidates in the combined entry generators. The improvement in F-score of 6-10 percentage points from the baseline models for the combined models is also significant in practice.

6 Discussion

In this section, we give a brief overview of previous and related work on analogy and machine learning of morphology. For a full review of analogy, see Hoffman (1995). We only review some of the central concepts and provide pointers to the use of analogy in phonology and morphology. The status of analogy as an explanatory device and learning mechanism in cognitive science is well established. We relate the concept of analogical learning to other well-established machine-learning concepts. With this broad backdrop, we wish to motivate how some machine-learning efforts on automated entry generation for morphologies are comparable to the method we propose even if they may not have been conceived as instances of analogical learning. Many of the early efforts on entry generation from the 1990s focused on entry generation for parsers and machine translation systems and not so much on entries for morphological analyzers. A more in-depth comparison with these efforts is a topic for future work.

In 6.1, analogy as a cognitive concept is presented. In 6.2, we introduce some key aspects of machine learning of morphology. In 6.3, we present some previous efforts in using analogical machine learning for natural language processing. In 6.4, we discuss methods related to our approach. In 6.5, we compare test results with previous efforts. In 6.6, we give some notes on the implementation of the methods. In 6.7, we discuss some future work.

6.1 Analogy as a Cognitive Concept

We first look at analogy as a general cognitive concept, and then we turn to how analogy as a cognitive concept influences linguistic thinking. Hoffman (1995) provides a survey of the concepts covered by analogy from the early Greeks to modern computer science. He points out that the usage of analogy has evolved from being a concept of geometric proportions, A is to B as C is to D, to a creative rhetorical device, where A is in B as C is in D, and then finally to something that can be seen as a fundamental cognitive process. Hoffman deconstructs the concept and demonstrates that what we perceive as an analogy often depends on the inference constraints we agree on or are able to defend.

Current efforts in cognitive research (Kurtz and Loewenstein, 2007; Gentner & al, 2004; Loewenstein & al, 2003) have shown that persons who are told to find the structural similarities between two seemingly unrelated problems are much more likely to use this structure when solving new problems or for transferring solutions from old problems than persons who are not told to relate the training problems. This has several implications, the least controversial of which is that humans seem to apply analogical reasoning better after some minimally supervised training.

In linguistics, Itkonen and Haukioja (1997) show how analogies which hold both on the level of form, as well as on the level of meaning can be computed by relying at the same time on the surface and the structural representations. The generative approach to linguistics maintains that all surface forms are generated from abstract underlying forms, except maybe for a few odd cases. However, the pre-generative idea that surface forms can influence other surface forms, (e.g. paradigmatic analogy in historical linguistics) has reemerged in a number of formal models; e.g. Eddington (2006) demonstrates that analogy can predict all surface instances of words, not just the exceptional cases. Keuleers & al (2007) develop the view that word inflection is driven partly by non-phonological analogy, e.g. with orthography, semantics, etc., and that non-phonological information is of particular importance to the inflection of non-canonical roots, i.e. the inflection of new words. In two experiments, they demonstrate that analogy as a process is the most likely explanation for human inflection of new and previously unseen words and that this can be modeled in a computer simulation.

6.2 Analogy and Machine Learning of Morphology

We review some of the fundamental concepts underpinning machine learning and its application to morphology learning. Finally, we characterize analogy in terms of current machine learning concepts. Initially machine learning was concerned with whether negative evidence is needed and what can be learned from positive evidence alone. Early on, Gold (1967) showed that context-sensitive languages require an informant, i.e. both pre-tagged positive and negative information, and that not even regular languages can be learned from text alone, i.e. from raw positive information. That formal languages cannot be learned without negative evidence and that negative evidence is not readily available to children are two facts that have been widely used as evidence that learning language is special, i.e. the basic building blocks of language are largely innate. This line of reasoning is known as the argument from the poverty of the stimulus. A range of evidence which challenges this line of argumentation has been put forward by Manning in (Bod & al, 2003), the most important of which is evidence by Horning (1969) that, contrary to categorical grammars, probabilistic grammars are learnable form positive evidence alone.

More recently the distinction between machine learning algorithms has been whether they learn in a supervised or an unsupervised fashion from positive evidence, i.e. do they learn from pre-tagged or raw input. Below, we review some of the supervised and unsupervised approaches to learning morphology. For an overview of recent work in the field of supervised and unsupervised learning of morphology, see Wicentowski (2002). For an overview of work in unsupervised segmentation and learning of morphology, see Goldsmith (2008).

At the extreme of the supervised spectrum of morphology learning algorithms, we have e.g. Murf created by Carlson (2005). Murf is a program intended to induce a morphological transducer from traditional-style hand-tagged inflectional paradigm sets as training data augmented with negative evidence for exceptions. From a linguistic point of view Murf can be seen to implement some of the traditional principles of taxonomical morphemic analysis. Another system by Oflazer & al (2001) uses a semiautomatic technique for developing finite-state morphological analyzers for use in natural language processing applications. The system generates finite-state analyzers from information elicited from human informants. Their approach uses transformation-based learning to induce morphographemic rules from examples and combines these rules with the elicited lexicon information to compile the morphological analyzer. As they themselves point out, there are also other opportunities for using machine learning in the acquisition process. For instance, one of the important issues in wholesale acquisition of open-class items is that of determining which paradigm a given citation form belongs to.

In order to achieve a morphological labeling of paradigmatic segments, we need hand-tagged data in the form of paradigms and exemplars to learn from, i.e. we need supervised learning, but the challenge is to produce at least some of the training material for supervised algorithms in an unsupervised way in order to achieve minimal supervision. E.g. Yarowsky and Wicentowski (2000) use monolingual context such as word vectors and the Levenshtein distance between words to find morphologically corresponding verb forms, and Yarowsky & al (2001) use a multilingual context in the form of word correspondences in bilingual corpora to identify likely base forms for inflected verb forms with high accuracy. It had been noticed by Brown et al. (1992) that words that are semantically related have a higher probability than chance to co-occur within a window of 3 to 500 words of each other in running text, which led Baroni & al (2002) to use mutual information between pairs of nearby words in a corpus as a crude measure of semantic relatedness, whereas Moreau & al (2007) use documents in the same domain to find morphologically corresponding word forms.

At the other extreme in the unsupervised end of the spectrum, is morphology learning from unstructured text without any supervision, i.e. methods which identify likely morphs and their morphotax. For language modeling, it is useful to have an algorithm which compresses a corpus in such a way as to generate a description of the corpus and the lexicon which is as brief as possible, also known as the general principle of Minimum Description Length. It turns out that this is directly useful in speech recognition (Creutz & al., 2006) as the language model can be optimized to contain more elaborate co-occurrence estimates for the most likely segments and their combinations while relying on back-off estimates for the remaining segments. Using unsupervised methods, it is also possible to identify segments which could form paradigms (Goldsmith, 2007). The goal of the unsupervised morphology discovery methods is not to produce a labeling, which is a human interpretation of the morph-classes that may have been discovered, but to discover a morph inventory and to provide an account for how these morphs co-occur.

In view of the machine learning concepts mentioned above, when defining analogy as a proportional relation between words “A : B :: C : D”, we need to specify at least the words A, B and C in order to infer D, which easily casts analogy as a supervised learning method. However, if in some unsupervised way, as outlined above, we can identify the words A, B and C that are likely to be semantically related in a proportional relation, analogy becomes a completely unsupervised or at most a minimally supervised method.

6.3 Analogy in Natural Language Processing

Analogy has been used as a device to explain phenomena on all levels of natural language, i.e. phonology, syntax and semantics, and to process various kinds of representations used in linguistic applications. On the phonological level, Goldsmith (2007) uses analogy mostly as an explanatory device for his algorithm for morpheme segmentation, Linguistica, which aims at discovering inflectional paradigms. He points out that an analogy like: *charge : change :: large : ?* is real but not particularly useful and that the frequency of occurrence can be used for identifying more useful analogies like: *charge : charged : grade : graded*. Goldsmith (2008) notes that analogy has the advantage that it can say something useful even about generalizations that involve a very small number of pairs (e.g., *say : said :: pay : paid*). It is more difficult for a purely unsupervised approach to become aware that those pairs of words should be related.

In syntax, Itkonen and Haukioja (1997) studied analogies as a relation between surface and structure. In lexical semantics, Turney (2008) observes that recognizing analogies, synonyms, antonyms, and associations appear to be four distinct tasks, requiring distinct NLP algorithms. In the past, the four tasks have been treated independently, using a wide variety of algorithms. He suggests a supervised corpus-based machine learning algorithm for

classifying analogous word pairs, and shows that it can solve multiple-choice SAT analogy questions, TOEFL synonym questions, ESL synonym-antonym questions, and similar-associated-both questions from cognitive psychology. He argues that analogy provides a framework that has the potential to unify the field of semantics, if we subsume synonyms, antonyms, and associations under analogies. He points out that, in essence, X and Y are antonyms when the pair "X : Y" is analogous to the pair "black : white", X and Y are synonyms when they are analogous to the pair "levied : imposed", and X and Y are associated when they are analogous to the pair "doctor : hospital".

In what can be characterized as a surface approach to compositional semantics, Lepage and Denoual (2005) present a machine translation system between English, Japanese, Chinese and Korean and demonstrate that the only thing needed by machine translation between these languages is analogy and a sufficiently large bilingual training corpus. They rely on the aligned sentences of the training corpus to be semantically equivalent.

Stroppa and Yvon (2006) generalize the notion of analogical proportions to a generic algebraic framework capable of handling a number of representations that are commonly encountered in linguistic applications from raw text to syntax trees, e.g. words over a finite alphabet, feature structures, labeled trees, etc. For each case, they provide algorithms and discuss related work.

6.4 Analogy and Exemplar-based Learning Methods

Most exemplar-based learning methods can subscribe to the idea that for an exemplar A there is a pre-labeled or transformed exemplar B, and the idea is that for a new and previously unseen instance C, we wish to predict a labeled or transformed instance D. The difference between the methods is mostly in how directly they rely on the training material for making the predictions. Research shows that a larger set of training exemplars "A : B" gives better predictions. Common sense has it that a larger set of training exemplars will take longer to process, so to speed-up the prediction process, we need to preprocess the training data: either we reduce the data to the most essential and distinctive exemplars or we extract the most likely labeling or transformation rules.

Memory-Based Learning (MBL) is applicable to a wide range of tasks in Natural Language Processing (NLP). Memory-Based Learning is a direct descendant of the classical k-Nearest Neighbor (k-NN) approach to classification, which makes predictions based on the pre-labeled k most similar neighbors from which an outcome distribution can be calculated. Daelemans (2007) has been working since the end of the 1980s on the development of Memory-Based Learning techniques and algorithms. The Analogical Modeling algorithm by Skousen (1989) differs from k-NN in that it groups neighbors according to similarity of context and consistency of prediction. This may give more emphasis to distant groups of neighbors with consistent predictions compared to k-NN. A consistent prediction for a context may contribute more than one outcome but the algorithm gives preference to contexts which contribute a single outcome. Skousen's algorithm works directly on the data and retrieves all the analogies in order to estimate a probability distribution of the outcome. For a brief overview, see Skousen (2003).

Lepage (1998) directly considers all combinations of A and B in the training data for making predictions, but the similarity of A to B and A to C is guided by the edit distance and the fact that all segments of A and D must be accounted for either by B or by C. See Figure 6 for an illustration. Yvon (2003) generalizes the notion of analogy between strings introduced by Lepage by providing a definition of the notion of analogical proportion between strings "A : B :: C : D" over a finite alphabet. Given the definition, Yvon demonstrates that solving an analogical equation, i.e. finding the fourth term of proportion can be performed using finite-state transducers. He uses the degree of analogy, i.e. the number of discontinuous

segments in the analogy, as the weighting scheme arguing that fewer is better, which can be used when reducing the search space. In addition, the frequency of the various outcomes generates a prediction distribution.

$$\begin{array}{cccc} A_1 & A_2 & A_3 & \\ & & & :: \\ B_5 & B_2 & B_6 & \end{array} \quad \begin{array}{ccc} C_1 & C_2 & C_3 \\ & & \\ D_5 & D_2 & D_6 \end{array}$$

Fig. 6[7]. Identical segments of A and D in C and B have the same indexes.

Mikheev (1996, 1997) presents a technique for statistical acquisition of rules which guesses possible parts-of-speech for unknown words. A full-form lexicon with part-of-speech labels provides exemplars of inflected forms and pre-labeled base forms to learn rules from. A raw corpus provides frequencies of inflected forms for evaluating the positive and negative impact of the proposed rules. Three complementary sets of word-guessing rules are induced: prefix morphological rules, suffix morphological rules and ending-guessing rules. The learning is performed on the Brown Corpus data. The method relies on verifying some of the suggestions in a lexicon, which gives a good performance on medium frequency words, i.e. frequency rank 30,000-90,000, that tend to be words derived from the core vocabulary of the language, but not yet the most esoteric low-frequency strings.

In our work, we use the method presented by Lindén (2008a), which estimates the probability of the possible prefix and suffix transformations from A to B using, e.g. a full-form lexicon and a raw corpus. The probability of the core or stem segment copied from C to D decreases relative to the length of the segment. This fulfills the basic requirement on analogical proportions by Lepage that all segments of A and D must be accounted for either by B or by C as illustrated in Figure 6. The transformations are encoded as weighted finite-state transducers.

Lepage (2001) introduces the notion of concatenation of analogical operations, which Lindén (2008a) uses for decomposition in order to estimate the probabilities of analogies on prefixes, stems and suffixes separately. By pre-compiling the most frequent analogies found in the training data, we can assign probabilities to the analogy patterns in advance and use this pattern collection to speed up the calculations of the complete set of analogies and related probabilities. We argue that more likely transformations are better, and that transformations where A covers as much of C as possible are better, as such transformations represent attested information.

Skousen's (1989) algorithm treats the positions in a string as independent variables and all combinations are considered, i.e. also the non-consecutive ones. Lindén's (2008a) motivation for considering only consecutive positions is that inflectional information tends to be at the extremes of a word or conditioned on the extremes of the root to which it is affixed or infixed. This tends to generate long suffixes and prefixes for the analogies potentially requiring more training data, but it also allows for very accurate matching when the training data is available as demonstrated in Lindén (2008b).

Lepage (2000) proves some theorems characterizing the generative power of languages on analogical strings. He shows e.g. how reduplication, *aa*, and bounded center embeddings, $a^n b^m c^n$, can be explained via repeated application of the analogies *a : aa*, as well as *abc : abbc* and *abc : aabcc*. As demonstrated by Lindén (2008), this is more power than currently needed to cover most of the active morphological phenomena in new and previously unseen words of English, Finnish, Swahili and Swedish. Arguably the

[7] Figure 6 is intended only as an illustration of the basic requirements of analogy. An analogy may be considerably more complex, e.g. when translating from one language to another there may be many more segments and the segments may be reordered, but all the segments in the source and the analogue target still need to be accounted for in the target or the analogue source.

reduplication of verb stems in past tense in Swahili could use a more powerful pattern induction mechanism, but generally this is not needed because new verbs occurring in past tense will also very likely occur in e.g. the present tense, so there are several opportunities to learn a new verb in order to identify its paradigm and base form. More power is probably required only in higher-level NLP tasks, e.g. syntactic analysis and translation as introduced by Lepage and Denoual (2005).

6.5 Comparison with Results from Similar or Related Efforts

Similar efforts have been made on other sets of test data and some insights can be gleaned from comparing with them, even if a direct comparison is difficult.

Stroppa and Yvon (2005) present experimental results obtained on a morphological analysis task guessing base form and morphological features for an inflected form in English with the following precision and recall: nouns 75 % precision and 95 % recall; verbs 95 % and 97 %; adjectives 28 % and 88 %, respectively. It is interesting to note that verb forms are the easiest to get right, whereas it is much trickier to guess the correct base forms and syntactic features of nouns and adjectives. The explanation is probably that the base forms of nouns and adjectives have much more varied character patterns, so there will be candidates suggesting analogies both with and without inflectional endings for many strings, whereas verb endings tend to be more easily identified.

Wicentowski (2004) presents the WordFrame model, a noise-robust supervised algorithm capable of inducing morphological analyses for languages which exhibit prefixation, suffixation, and internal vowel shifts. In combination with a naive approach to suffix-based morphology, this algorithm is shown to be remarkably effective across a broad range of languages, including those exhibiting infixation and partial reduplication. Results are presented for over 30 languages with a median accuracy of 97.5 % on test sets including both regular and irregular verbal inflections. The excellent accuracy is partly explained by the fact that he uses a dictionary to filter the suggested base forms. His results are very good, but should be seen in the light of the results by Yvon and Stroppa (2005), where a substantial challenge seems to be in modeling the behavior of nouns and adjectives, which are also the most frequent categories among new words.

Claveau and L'Homme (2005) label morphologically related words with their semantic relations using morphological prefix and suffix analogies learned from a sample of pre-labeled words with a recall of 72 % and precision of 65 % on separate test data.

Baldwin (2005) acquires affix and prefix transformations achieving 0.6 F-score for English using Timbl as the classifier, but the classification was for syntactic features not for inflectional paradigm.

We recall that our model is developed for guessing the paradigms of unknown and previously unseen inflected words, i.e. their base forms cannot be tested against a lexicon. In view of the results from comparable reports from other languages, our results as shown in Table 8 for the combined model are very good, because the data shows that the final entry generators have 77-81 % precision and 89-97 % recall, i.e. an F-score of 84-88 %, on languages with different morphological complexity. It is perhaps to be expected that our model has the lowest recall for Finnish, which is morphologically the most complex of the three test languages. Due to the way the Swedish lexicon is constructed, it is also very plausible that precision is lowest for the Swedish entry generator as the set of paradigms was very fine-grained.

Even a fairly loose but extensive lexical model is able to improve the performance of a purely statistical corpus-based model significantly, and vice versa. It should be noted that one reason for the low performance of the corpus-based model, is that the probabilities were estimated separately for the transformations from word form to base form and for base form

to paradigm. The advantage is that we need less training data for each of the two stages, but it disconnects some of the constraints between the inflected forms and the paradigms in the corpus-based model. However, the lexical model brings back the connections between an inflected form and its paradigm, which shows in the performance of the combined model. We can regard the corpus-model as encoding the performance of a human language model as manifested in corpora, and the lexicon model as generally encoding the competence of a native speaker. It is likely that the combination of the corpus information with e.g. stricter hand-made lexical models might further benefit the outcome by giving priority to more frequent sound patterns and paradigms that are known to be productive. What paradigms are perceived to be productive is, however, a task that is dependent on the size of the existing lexicon that is to be extended and the frequency range from which the new words are drawn.

When studying the words for which the hybrid classifier failed to generate anything at all, we notice that they were uppercase strings for which the hand-made and the statistical model generated different suggestions. When put together, the statistical and hand-made generator had no common suggestion for these in the hybrid model. Converting the strings to lowercase is not a catch-all option, because many of them are acronyms for which case is significant. Acronyms are sometimes inflected according to the regular words, which they are abbreviations of, and sometimes according to their phonology as acronyms.

A quick look at the words which fail for English reveals that among them are e.g. *preacheth, Surmountheth, corruptehth,* which could not have received a correct guess as out-dated verb forms were not available as analogical models. We require the correct answer to have a specific base form and a paradigm which indicates all the correct inflected forms. E.g. the word *equalizes* gets the base form *equaliz*, whereas we expect *equalize*, even if the word is otherwise correctly identified as a verb. We also require that words like *plowman* are correctly identified as having the plural *plowmen.* It is not enough just to identify it is as some noun. The same goes for other words with irregular forms or deficient paradigms. This illustrates how entry guessing is more difficult than guessing the part-of-speech: a correct lexical entry requires guessing a base form and a paradigm that together produce all the correct forms and only the correct forms for an out-of-vocabulary word.

An aspect that is not considered in this work is the fact that a word may appear in a corpus in several forms which together support one or more lexical entries, i.e. base forms with paradigm information. Forsberg et al. (2006) used an approach that automatically deduces extraction rules for which they could find as much support as was logically possible in order to make a safe inference. This leads to rules safely extracting words that already have a number of word forms in the corpus, i.e. mid or high-frequency words. Such methods are especially suitable for resource-poor languages lacking readily available public domain morphological descriptions like the Ispell dictionaries (Kuenning, 2007) or similar. Forsberg et al. (2006) concluded that it is recommendable that a linguist writes the extraction rules. Lindén and Tuovila (2009a, 2009b) look for inflected word forms of each base form and paradigm combination to determine which lexical entries are best supported in a given corpus. However, it turns out to be difficult to gain significant improvements over the method presented in this article by using additional word forms, as new or infrequent words by definition appear only in one or two forms. These forms do not necessarily distinguish between very fine-grained paradigm descriptions. In such cases, more data or the evidence from a native speaker is still necessary to finalize the selection between the top candidates.

We did a follow-up experiment using the entry generator for adding new words to our Finnish lexicon. We generated key word forms for the top 6 base form and paradigm candidates from which a native speaker selected the correct lexical entries. It turned out that the revising speed for a native speaker is 300-400 words per hour (Listenmaa, 2009). Consequently, adding e.g. 60 000 new entries to a lexicon using the proposed method can be achieved by 10 native speakers working in parallel in a few days.

6.6 Implementation note

The models have been implemented with a cascade of weighted finite-state transducers. For conveniently creating morphological analyzers and entry generators, *HFST–Helsinki Finite-State Technology* (HFST, 2008; Lindén & al, 2009) is available as an Open Source toolkit. Open Source tools for general manipulation of weighted finite-state transducers have been implemented by, e.g., Allauzen & al (2008) and Lombardy & al (2004).

The entry guesser transducers correspond to what Lepage (2001) characterizes as the immediate analogical derivation of pure analogy, i.e. no reduplication is modeled, which is not such a big loss in practice, cf. Section 6.4. We also do not compute the transitive closure of the lexical analogies, which makes the derivation process relatively fast as it requires only a fixed number of composition operations. Instead we need a fairly large set of pre-compiled analogy patterns in the transducer.

6.7 Future work

In the same way that Lindén (2008) determined that 5000 words seems to be enough as training material for a base form guesser, it might also be interesting to see how small lexicons the entry generator is able to generate useful results from, while still speeding up the human post-processing. This is relevant for resource-poor languages building morphological analyzers from scratch.

The current approach only considers information available within a word, even if the context of a new word also contributes to the lexical information human readers infer. When representing context information, a unification-based approach was predominant in the 1990's. How context can assist in inferring information for lexical entries has been studied e.g. by Barg and Kilbury (2000) and Barg and Walther (1998) in the project Dynalex–Dynamic extension of the Lexicon. Earlier efforts focused on extracting lexical context by parsing typeset dictionaries into machine-readable form, which was explored e.g. in the Acquilex project (Copestake, 1992). Representing complex contexts using finite-state transducers is becoming increasingly possible and opens up interesting research aspects for lexical acquisition for finite-state parsers and translation engines.

7. Conclusion

We tested our models for classifying inflected forms of new words by analogy with a set of lexical entries from three different languages types. We tested on Finnish, which is a highly inflecting Finno-Ugric language with a considerable set of stem change categories and multi-stem compounding, as well as on Swedish, which is a Germanic language with a fair amount of regular and irregular inflectional patterns and a multi-stem compounding mechanism. For comparison we also tested on English which is a Germanic language known to have very little regular inflectional morphology but a reasonable set of irregular and deficient inflectional paradigms with a very restricted multi-stem compounding. Our hybrid model achieved 77-81 % precision and 89-97 % recall, i.e. an F-score of 84-88 %. The average position for the first correctly generated entry was 1.6-2.0. A study demonstrated that a native speaker can revise suggestions from the morphological entry generator at a speed of 300-400 entries per hour.

Acknowledgements

I am grateful to the Finnish Academy and to the Finnish Ministry of Education for funding the project. I am also grateful to Tommi Pirinen, Jussi Tuovila and Anssi Yli-Jyrä for many fruitful discussions as well as to Lars Borin, Yves Lepage and Heiki-Jaan Kaalep for valuable comments on the manuscript.

References

Allauzen, Cyril, Michael Riley, Johan Schalkwyk, Wojciech Skut and Mehryar Mohri. 2007. OpenFst: A General and Efficient Weighted Finite-State Transducer Library, *Lecture Notes in Computer Science,* pages 11-23.

Barg, Petra, and James Kilbury. 2000. Incremental Identification of Inflectional Types. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 49 - 54. Saarbrücken, Germany

Barg, Petra, and Markus Walther. 1998. Processing unknown words in HPSG. In *ACL-36: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. pages 91— 95.

Bod, Rens, Jennifer Hay and Stefanie Jannedy (eds.). 2003. *Probabilistic Linguistics.* MIT Press.

Baldwin, Timothy. 2005. Bootstrapping Deep Lexical Resources: Resources for Courses. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, Association for Computational Linguistics, pages 67-76.

Baroni, Marco, Johannes Matiasek and Harald Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning, SIGPHON-ACL*, pages 11-20.

Brown, Peter. F., Peter V. deSouza, Rober L. Mercer, Vincent J. Della Pietra and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18:467-479.

Carlson, Lauri. 2005. Inducing a Morphological Transducer from Inflectional Paradigms. In *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday,* CLSI Publications, ISSN 1557-5772, Stanford University, pages 18-24.

Claveau, Vincent, and Marie-Claude L'Homme. 2005. Structuring Terminology using Analogy-Based Machine Learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, Copenhagen, Denmark, pages 17-18, August.

Copestake, Ann. 1992. The ACQUILEX LKB: Representation Issues in Semi-Automatic Acquisition of Large Lexicons. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92*), pages 88 – 96.

Creutz, Mahtias., Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based Speech Recognition and Modeling of Out-of-Vocabulary Words across Languages. In *ACM Transactions on Speech and Language Processing*, Vol. 5, No. 1, Article 3, 29 pages

Creutz, Mahtias, Krista Lagus, Krister Lindén and Sami Virpioja. 2005. Morfessor and Hutmegs: Unsupervised Morpheme Segmentation for Highly-Inflecting and Compounding Languages. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, Tallinn, Estonia.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch. 2007. TiMBL: Tilburg Memory-Based Learner, version 6.1, Reference Guide. *Technical Report–ILK 07-07*, Department of Communication and Information Sciences, Tilburg University, 64 pages.

Eddington, David. 2006. Paradigm Uniformity and Analogy: The Capitalistic versus Militaristic Debate. *IJES, International Journal of English Studies*, 6 (2): 1-18.

Forsberg, Markus, Harald Hammarström and Aarne Ranta. 2006. Morphological Lexicon Extraction from Raw Text Data. FinTAL 2006, LNCS 4139, pages 488-499.

FreeLing 2.1–An Open Source Suite of Language Analyzers (computer file). 2007.
Available at: http://garraf.epsevg.upc.es/freeling/

Gentner, Dedre, Jeffrey Loewenstein, Leigh Thompson. 2004. Analogical Encoding: Facilitating Knowledge Transfer and Integration. In K. Forbus, D. Gentner, & T. Regier (Eds), *Proceedings of the 26th Meeting of the Cognitive Science Society,* pages 452-457.

Gold, E. Mark. 1967. Language Identification in the Limit. *Information and Control*, 10(5):447-474.

Goldsmith, John. 2007. *Morphological Analogy: Only a Beginning*.
Available at: http://hum.uchicago.edu/~jagoldsm/Papers/analogy.pdf

Goldsmith, John. 2008. *Segmentation and morphology*. Departments of Linguistics and Computer Science, The University of Chicago. (To appear in The Handbook of Computational Linguistics).
Available at: http://hum.uchicago.edu/~jagoldsm//Papers/segmentation.pdf

HFST – Helsinki Finite-State Technology (computer file). 2008.
Available at: http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/index.shtml

Hoffman, Robert R. 1995. Monster Analogies. *Artificial Intelligence Magazine*, 16(3):11-35.

Horning, James Jay. 1969. *A Study of Grammatical Inference*. PhD Thesis. Stanford University.

Itkonen, Esa, and Jussi Haukioja. 1997. A rehabilitation of analogy in syntax (and elsewhere). In András Kertész (ed.), *Metalinguistik im Wandel: die cognitive Wende in Wissenschafttheorie und Lingusitik*. Frankfurt am Main. Peter Lang, pages 131-177.

Keuleers, Emmanuel, Dominiek Sandra, Walter Daelemans, Steven Gillis, Gert Durieux and Evelyn Martens. 2007. Dutch Plural Inflection: The Exception That Proves the Analogy. *Cognitive Psychology*, 54(4), pages 283-318.

Koskenniemi, Kimmo. 1983. Two-Level Morphology: A General Computational Model of Word-Form Recognition and Production. Publications No 11. University of Helsinki, Department of General Linguistics.

Kuenning, Geoff. 2007. *Dictionaries for International Ispell*.
Available at: http://www.lasr.cs.ucla.edu/geoff/ispell-dictionaries.html

Kurimo, Mikko, Mathias Creutz and Ville Turunen. 2007. Overview of Morpho Challenge in CLEF 2007. *Working Notes of the CLEF 2007 Workshop*, pages. 19-21.

Kurtz, Kennth J., and Jeffrey Loewenstein. 2007. Converging on a New Role for Analogy in Problem Solving and Retrieval. In *Memory & Cognition* 35(2):334-341.

Lepage, Yves. 1998. Solving Analogies on Words: An Algorithm. COLING-ACL, pages 728-734.

Lepage, Yves. 2000. Languages of Analogical Strings. In *Proceedings of the 18th conference on Computational linguistics,* vol. 1, pages 488 – 494. Saarbrücken, Germany.

Lepage, Yves. 2001. Analogy and Formal Languages. In *Proceedings of FG/MOL 2001*, pages 373-378.

Lepage, Yves, and Etienne Denoual. 2005. Purest ever Example-based Machine Translation: Detailed Presentation and Assessment'. In *Journal of Machine Translation* 19, pages 251–282.

Lingsoft, Inc. 2007. *Demos.*
Available at: http://www.lingsoft.fi/?doc_id=107&lang=en

Lindén, Krister. 2006. Multilingual Modeling of Cross-lingual Spelling Variants. In *Journal of Information Retrieval*, vol 9, pages 295-310.

Lindén, Krister. 2008a. A Probabilistic Model for Guessing Base Forms of New Words by Analogy**.** In *CICling-2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics*, Haifa, Israel.

Lindén, Krister. 2008b. Assigning an Inflectional Paradigm using the Longest Matching Affix**.** In *Ei mitään ongelmia. Juhlakirja* Juhani Reimanille *50-vuotispäiväksi 23.1.2008* Eds. Matti Wiberg and Antti Koura. Turku 2008.

Lindén, Krister. 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.

Lindén, Krister, and Jussi Tuovila. 2009a. Corpus-based Paradigm Selection for Morphological Entries. In *Proceedings of NODALIDA 2009*, May, Odense, Denmark.

Lindén, Krister, and Jussi Tuovila. 2009b. Corpus-based Lexeme Ranking for Morphological Guessers. In *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology 2009*. September, Zürich, Switzerland.

Lindén, Krister, Miikka Silfverberg and Tommi Pirinen. 2009. HFST Tools for Morphology– An Efficient Open-Source Package for Construction of Morphological Analyzers. In *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology 2009*. September, Zürich, Switzerland.

Listenmaa, Inari. 2009. Combining Word Lists: Nykysuomen sanalista, Joukahainen-sanasto and Käänteissanakirja (in Finnish). *Bachelor's Thesis*. Department of Linguistics, University of Helsinki.

Loewenstein, Jeffrey, Leigh Thompson and Dedre Gentner. 2003. Analogical Learning in Negotiation Teams: Comparing Cases Promotes Learning and Transfer. *Academy of Management Learning and Education*, 2(2):119-127.

Lombardy, Sylvain, Yann Régis-Gianas and Jaques Sakarovitch. 2004. Introducing Vaucanson**.** *Theoretical Computer Science,* 328(1-2):77 – 96.

Mikheev, Andrei. 1996. Unsupervised Learning of Word-Category Guessing Rules. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, pages 327-334.

Mikheev, Andrei. 1997. Automatic Rule Induction for Unknown-Word Guessing. In *Computational Linguistics*, 23(3):405-423.

Moreau, Fabienne, Vincent Claveau and Pascale Sebillot. 2007. Automatic Morphological Query Expansion Using Analogy-Based Machine Learning. *Advances in Information Retrieval, Lecture Notes in Computer Science*, pages 222-233.

Nykysuomen sanalista (computer file). 2007.
Available at: http://kaino.kotus.fi/sanat/ nykysuomi/

Oflazer, Kemal, Sergei Nirenburg and Marjorie McShane. 2001. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. In *Computational Linguistics* 27(1):59-85.

Pirinen, Tommi. 2008. Open Source Morphology for Finnish using Finite-State Methods (in Finnish). *Master's Thesis*. Department of Linguistics, University of Helsinki.

Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Kluwer.

Skousen, Royal. 2003. Analogical Modeling: Exemplars, Rules, and Quantum Computing. Presented at the *Berkeley Linguistics Society Conference.*

Stroppa, Nicolas, and François Yvon. 2005. An Analogical Learner for Morphological Analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*, pages 120–127, Ann Arbor, June.

Stroppa, Nicolas, and François Yvon. 2006. Formal models of analogical proportions. Technical Report D008. *Télécom Paris D*, ISSN: 0751-1345, Telecom ParisTech - École Nationale Supérieure de Télécommunications.

Turney, Peter. 2008. A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), August, Manchester, UK, pages 905-912.

Westerberg, Tom. 2008. *Den stora svenska ordlistan* (computer file)
Available at: http://www.dsso.se/

Wicentowski, Richard. 2002. Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework. *PhD Thesis*. Baltimore, USA.

Wicentowski, Richard. 2004. Multilingual Noise-Robust Supervised Morphological Analysis using the WordFrame Model. In *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology, ACL*, pages 70-77.

Yarowsky, David, and Richard Wicentowski. 2000. Minimally Supervised Morphological Analysis by Multimodal Alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Yarowsky, David, Grace Ngai and Richard Wicentowski. 2001. Inducing Multi-lingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *HLT '01: Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, pages 1-8.

Yvon, François. 2003. Finite-State Transducers Solving Analogies on Words, Technical Report D008. In *Télécom Paris D*, ISSN: 0751-1345. TELECOM ParisTech.