Estimating frequencies of inflected forms using simple frequency lists

Filip Graliński

Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Umultowska 87, 61-614 Poznań, Poland
filipg@amu.edu.pl

Abstract

Good estimates of probabilities of inflected forms are very important in the processing of languages with rich inflection. In this paper, we discuss how to obtain estimates of frequencies of inflected forms from raw frequency lists, without a corpus. The techniques described in this paper were evaluated indirectly in the context of Polish-to-English machine translation: by discarding inflected forms of low estimated frequencies it was possible to significantly reduce the translation time with some improvement in the translation quality.

Keywords: lexica, inflected forms, machine translation, homonymy

1. Introduction

Good estimates of the frequencies (or probabilities) of inflected forms are highly valuable in natural language processing for languages with rich inflection (such as Polish). First, such data fit naturally into probabilistic frameworks, e.g. when inflectional interpretations are combined into larger syntactic units with probabilistic grammars.

Second, even if no probabilistic framework is used in language analysis, the estimated frequencies of inflected forms can be used to discard spurious homonyms, i.e. cases when a theoretically possible inflected form of an infrequent lexeme is identical to a much more probable inflected form of another lexeme. For instance, the frequent Polish conjunction *albo* (= *or*) is accidentally homonymous with the vocative singular of the rare noun *alba* (= *alb*/*alba*). Obviously, the *alba* interpretation is only theoretical and for practical reasons should be simply discarded. More examples of spurious homonyms in Polish are given in Table [1] For the discussion of homonymy in Polish, see (Świdziński et al., 2003).

A simple frequency list of word types (uninterpreted strings) rather than a frequency list of interpreted forms is usually available, as the former is fairly trivial to obtain from any corpus (assuming a tokeniser is available), simply by counting the strings. We are going to show how to generate automatically a frequency list of **inflected** (interpreted) forms from a **raw** frequency list. The frequency list of interpreted forms will contain information on e.g. how frequent is *powieść* as the nominative of the noun *powieść* (= *novel*), as the accusative of this noun and as the infinitive of the verb *powieść* (= *succeed*).

Naturally, frequencies of inflected forms can be easily derived from a manually disambiguated corpus, however such corpus might not be always available, might be too small or the tagset used during the manual tagging might be not compatible with the one assumed in the particular NLP system. Larger corpora can be taken into consideration if automatic disambiguation (i.e. a POS-tagger) is applied, but again a POS-tagger might be unavailable or its tagset might be incompatible with the given system. Anyway, it may be the case that the whole corpus is unavailable for copyright reasons, just a simple frequency list was made available.

In the sequel, the following assumptions are made:

- a lexicon of inflected forms is available – inflected forms are grouped into lexemes, each lexeme is assigned a base form and a part-of-speech tag and each inflected form is labeled with an inflectional tag,
- a frequency list of uninterpreted words is available, the list was derived from some corpus, but the corpus itself is unavailable.

The aim is to assign an estimated frequency to each inflected form. The estimates need not be integers.

The idea is to use the following assumption: the distribution of inflected forms of the same part of speech is, at least to some extent, similar, e.g. it is assumed that the percentage of vocative forms for one feminine noun will be similar to the percentage of vocative forms for another such noun. For instance, let us take into consideration the word *albo*, and assume that no knowledge about the meaning of this word is given, except for the two possible inflectional interpretations (listed in Table). With no other knowledge, we could start with evenly distributing the frequency of string *albo* ($f(albo)$) into both interpretations ($f(albo)/2$ for the conjunction and $f(albo)/2$ for the vocative form). In effect, however, the distribution of inflected forms for the noun *alba* will be anomalous – all (or nearly all) forms of this noun are vocatives which is highly unusual for a feminine noun. Consequently, the estimated frequency of the vocative interpretation should be lowered. Such calculations and amendments could be carried out for some number of iterations for all ambiguous words.

[1] The level of "spuriousness" varies among these examples, as the theoretical interpretations for *albo* and *muszy* are rather bizarre for Polish native speakers, while the improbable interpretations of *bez* and *można* might be from time to time the right ones in real texts.

word	natural interpretation	improbable interpretation
albo	conjunction *albo* (= *or*)	vocative singular of *alba* (= *alb*/*alba*)
bez	preposition *bez* (= *without*)	nominative singular of *bez* (= *common lilac*) accusative singular of *bez* genitive plural of *beza* (= *meringue*)
można	predicate *można* (= *it is possible to*)	feminine nominative singular of *możny* (= *puissant*)
musi	present tense, 3rd person sing of *musieć* (= *must*)	masculine personal nominative plural of *muszy* (= *fly-like*)
mają	present tense, 3rd person plural of *mieć* (= *have*)	present tense, 3rd person plural of *maić* (= *deck*)
maj	nominative/accusative singular of *maj* (= *May*)	imperative of *maić* (= *deck*)

Table 1: Examples of spurious homonyms in Polish, based on (Jassem et al., 1998) and (Graliński, 2000)

In this paper, we discuss how the information on frequency of inflected forms can be used in a specific Polish-to-English rule-based machine translation system.

In Section the iterative algorithm for counting the frequencies of inflected forms is presented. The language resources are presented in Section , while in Section the evaluation of results is given. In Section , we discuss how to identify anomalies in lexicons using the information about the frequencies of inflected forms.

2. Algorithm

All the lexemes are divided into classes. The class of lexeme λ will be denoted by $c(\lambda)$. Lexeme classes are parts of speech with some more fine-grained distinctions:

- nouns are divided according to their gender, singularia[2] and pluralia tantum are assigned to distinct lexeme classes as well,
- perfective and imperfective verbs form two distinct classes,
- gradable and non-gradable adjectives form two distinct classes, and the same applies to adverbs.

These distinctions are made because of significant differences in the distribution of inflected forms for subgroups of the same part of speech, e.g. imperfective verbs have active participles, while perfective do not have such forms as a rule.

Let us denote the frequency of a word w (treated as an uninterpreted string) by $f(w)$ and the number of possible inflectional interpretations for w – as $n(w)$. The set of all possible inflectional interpretations of w is $\{\phi_1(w), \ldots, \phi_{n(w)}(w)\}$, where each $\phi_i(w)$ is a triple (w, λ, μ), where λ is the lexeme and μ is the inflectional tag of form $\phi_i(w)$ (e.g. *vocative singular*). The lexeme of inflected form ϕ will be denoted by $l(\phi)$ and its inflectional tag – by $m(\phi)$. The (estimated) frequency of inflected form ϕ will be denoted by $f(\phi)$. The aim of the algorithm is to estimate $f(\phi)$ in such a way that

$$f(\phi_1(w)) + \cdots + f(\phi_{n(w)}(w)) = f(w) \qquad (*)$$

for each word w. Note that $f(w)$ is known from the raw frequency list.

In the initialisation step $f(w)$ is distributed evenly among $f(\phi_i(w))$, i.e.

$$f^0(\phi_i(w)) = f(w)/n(w)$$

for each $i \in \{1, \ldots, n(w)\}$; f^k is the frequency at iteration step k, $k = 0$ denotes the initialisation step.

In particular, in case of unambiguous words (i.e. $n(w) = 1$):

$$f^0(\phi_1(w)) = f(w)$$

and it will remain the same through all the iteration steps, i.e.:

$$f^k(\phi_1(w)) = f(w)$$

for all k, as this is the exact value, no estimation will be calculated. From now on, only ambiguous words ($n(w) > 1$) will be considered.

In each iteration step we start with estimating the frequency of lexemes using the frequencies of inflected forms estimated in the previous step, simply by summing the frequencies of inflected forms:

$$f^k(\lambda) = \sum_{l(\phi)=\lambda} f^{k-1}(\phi),$$

where $f^k(\lambda)$ denotes the frequency of lexeme λ estimated at step k.

Then the estimated distribution of inflected forms (or, rather, their tags) is calculated for each lexeme class. Let $\delta(\xi, \mu)$ denote the estimated probability of inflectional tag μ for lexeme class ξ, then:

$$\delta^k(\xi, \mu) = \frac{\sum_{c(l(\phi))=\xi, m(\phi)=\mu} f^{k-1}(\phi)}{\sum_{c(l(\phi))=\xi} f^{k-1}(\phi)}$$

Now the expected frequency of each form is calculated based on the frequency of its lexeme and the probability of its inflectional tag[3]:

[2] In fact, most of Polish nouns usually described as *singulare tantum*, e.g. *fizyka* (= *physics*), *hojność* (= *generosity*), do have plural forms, though they are much less probable than singular ones.

[3] The same inflected form might be expressed with more than one word, e.g. there are two alternative forms of nominative plural for Polish noun *postać* (= *form*/*figure*), namely *postaci* and *postacie*. In such cases $\bar{f}^k$ should be multiplied by an additional discount factor. For the sake of simplicity, these details are skipped in the description of the algorithm.

$$\bar{f}^k(\phi) = F^k(l(\phi), w) \times \frac{\delta^k(c(l(\phi)), m(\phi))}{1 - \Delta^k(l(\phi), w)}$$

where F^k and Δ^k are defined as:

$$F^k(\lambda, w) = f^k(\lambda) - \sum_{l(\phi_i(w))=\lambda} f^{k-1}(\phi_i(w))$$

$$\Delta^k(\lambda, w) = \sum_{l(\phi_i(w))=\lambda} \delta^k(c(\lambda), m(\phi_i(w)))$$

F^k and Δ^k are used instead of simply f^k and 0 because we do not want to take into account the frequencies of forms of the current word w.

Now divide the set of lexemes for the given word w (namely, the set $\{l(\phi_i(w)) \mid i = 1, \ldots, n(w)\}$) into two disjoint subsets:

- $L_1(w)$ – the set of *one-form lexemes*, where one-form lexeme is a lexeme whose all inflected forms are expressed with one word,
- $L_m(w)$ – the set of lexemes with more than one distinct form.

Note that it is more difficult to estimate the frequency of one-form lexemes as there are no other forms to base estimations on.

There are two cases to consider for the given w:

1) Some of the lexemes are *one-form lexemes*, i.e. $L_1(w)$ is not empty.[4] Then, if $l(\phi_i(w))$ is **not** a one-form lexeme, then:

$$f^k(\phi_i(w)) = \bar{f}^k(\phi_i(w))$$

The rest of the frequency "mass" is evenly distributed among the one-form lexemes and then divided within such lexemes according to the distribution of inflected forms, i.e.:

$$f^k(\phi_i(w)) = r^k(w) \times \frac{1}{|L_1(w)|} \times \delta^k(c(l(\phi_i(w))), m(\phi_i(w)))$$

if $l(\phi_i(w))$ is a one-form lexeme, where:

$$r^k(w) = f(w) - \sum_{l(\phi_i(w)) \in L_m(w)} f^k(\phi_i(w))$$

2) None of the lexemes $l(\phi_i(w))$ is a one-form lexeme. Then, in order for equation (*) to be fulfilled, the frequencies of inflected forms have to be re-estimated by re-scaling $\bar{f}$:

$$f^k(\phi_i(w)) = \frac{\bar{f}^k(\phi_i(w))}{\sum_{j=1}^{n(w)} \bar{f}^k(\phi_j(w))} \times f(w)$$

[4] $L_m(w)$ may or may not be empty.

Example. Let us consider the word *bez* (i.e. $w = bez$). Then $n(w) = 4$ (see Table):[5]

- $\phi_1(w) = (bez, bez_{\texttt{noun:masc}}, \texttt{nom} : \texttt{sg})$
- $\phi_2(w) = (bez, bez_{\texttt{noun:masc}}, \texttt{acc} : \texttt{sg})$
- $\phi_3(w) = (bez, beza_{\texttt{noun:fem}}, \texttt{gen} : \texttt{pl})$
- $\phi_4(w) = (bez, bez_{\texttt{prep}}, -)$[6]

The most probable interpretation for a native speaker of Polish is definitely $\phi_4(w)$.

If the frequency list described in Section is used then $f(w) = 91532$. In the initialisation step it is divided evenly between interpretations, i.e.:

$$f^0(\phi_i(w)) = \frac{91532}{4} = 22883$$

We start the first iteration step with estimating the frequencies of lexemes:

- $f^1(bez_{\texttt{noun:masc}}) = 46277$
- $f^1(beza_{\texttt{noun:fem}}) = 22910$
- $f^1(bez_{\texttt{prep}}) = 22883$

The frequencies of lexemes $bez_{\texttt{noun:masc}}$ and $beza_{\texttt{noun:fem}}$ are obviously overestimated just because of the accidental homonymy with the preposition *bez* (the very homonymy which is considered now). The other forms[7] of $bez_{\texttt{noun:masc}}$ and $beza_{\texttt{noun:fem}}$ contribute only 511 and 27 to, respectively, $f(bez_{\texttt{noun:masc}})$ and $f(beza_{\texttt{noun:fem}})$, in other words:

- $F^1(bez_{\texttt{noun:masc}}, bez) = 511$
- $F^1(beza_{\texttt{noun:fem}}, bez) = 27$

Finally, $f^1(bez_{\texttt{prep}})$ is simply equal to $f^0(\phi_4(w))$ as *bez* is the only inflected form of preposition *bez*.

Now the distribution of inflected forms for all the lexeme classes is calculated. The relevant values are:

- $\delta^1(\texttt{noun} : \texttt{masc}, \texttt{nom} : \texttt{sg}) = 0.181$
- $\delta^1(\texttt{noun} : \texttt{masc}, \texttt{nom} : \texttt{sg}) = 0.181$
- $\delta^1(\texttt{noun} : \texttt{fem}, \texttt{gen} : \texttt{pl}) = 0.106$
- $\delta^1(\texttt{prep}, -) = 1.000$

Now the expected frequencies of inflected form are determined:

- $\bar{f}^1(\phi_1(w)) = 511 \times \frac{0.181}{1-(0.181+0.181)} = 144.970$
- $\bar{f}^1(\phi_2(w)) = 144.970$
- $\bar{f}^1(\phi_3(w)) = 27 \times \frac{0.106}{1-0.106} = 3.201$

[5] `masc` – masculine, `nom` – nominative, `sg` – singular, `acc` – accusative, `fem` – feminine, `gen` – genitive, `pl` – plural, `prep` – preposition

[6] No inflection tag is needed for a preposition.

[7] All the other forms of $bez_{\texttt{noun:masc}}$ and $beza_{\texttt{noun:fem}}$ are unambiguous as far as the lemmata are concerned.

($\bar{f}^1(\phi_4(w))$ is not calculated as $bez_{\texttt{prep}}$ is a one-form lexeme).

As one of the possible lexemes is a one-form lexeme, we follow case (1):

- $f^1(\phi_1(w)) = \bar{f}^1(\phi_1(w)) = 144.970$
- $f^1(\phi_2(w)) = \bar{f}^1(\phi_2(w)) = 144.970$
- $f^1(\phi_3(w)) = \bar{f}^1(\phi_3(w)) = 3.201$
- $f^1(\phi_4(w)) = f(w) - (f^1(\phi_1(w)) + f^1(\phi_2(w)) + f^1(\phi_3(w))) = 91238.859$

Just one iteration was enough to make the estimations of frequencies of inflected forms much more realistic.

3. Resources

The estimations of frequencies of Polish inflected forms were done using an extensive lexicon of Polish-to-English machine translation system Translatica (Jassem, 2006). The lexicon contains 87618 Polish lexemes with 1593440 inflected forms.

The frequency list derived from *PWN Corpus of Polish*[8] (including the archives of *Rzeczpospolita* newspaper) was used. It should be emphasised that the corpus itself was unavailable for legal reasons, only the frequency list was disclosed. The total number of words listed in the frequency list was 1200712. The size of the corpus was 91973930 words.

In fact, the frequency list contained the results of some simple normalisation function rather than the words themselves. The normalisation consisted in:

- converting to lower case,
- truncating a word after 20 letters and adding + character,
- removing periods (in case of abbreviations ending with a period).

For instance, the word *Parangaricutirimicuaro* would be normalised into *parangaricutirimicua+*. It is not difficult to take such normalisation function into account in the algorithm described in Section – the set of all inflected forms of words for which the normalisation function returns w should be simply considered.

4. Evaluation

We decided to apply an indirect evaluation, consisting in checking the results of machine translation using the lexicon with entries discarded according to the estimated frequencies. This approach was motivated by two reasons:

- the algorithm described in Section was developed while trying to improve the quality of machine translation,
- no disambiguated corpus with tagset compatible with the Translatica lexicon was available.

[8] `http://korpus.pwn.pl/index_en.php`

The experiment involved running Translatica system with various versions of the lexicon of inflected forms.

For the evaluation, a sample subcorpus (64061 segments) of Europarl English-Polish parallel corpus was used (Koehn, 2005). Translations were automatically evaluated using Meteor evaluation system (version 1.3 with standard settings for English) (Denkowski and Lavie, 2011).

The results of evaluation are presented in Table All the translations were done using Translatica – a traditional rule-based Polish-English machine translation system. Neither POS tagger nor target language model is used in Translatica.

The four versions presented in Table differ only in the lexicon of inflected forms. ALL FORMS is the baseline – the full lexicon with all the inflected forms, including spurious homonyms. In MANUALLY DISCARDED the inflected forms marked manually or semi-manually as unwanted in an *ad hoc* manner during the years of development of Translatica were removed from the lexicon. In AUTOMATICALLY DISCARDED the removal of inflected forms was based on the estimations of form frequencies calculated using the algorithm described in Section Namely, a form $\phi_i(w)$ is discarded if there exists form $\phi_j(w)$ for some $j \neq i$ such that $l(\phi_i(w)) \neq l(\phi_j(w))$ and $\theta f(\phi_i(w)) < f(\phi_j(w))$ for threshold θ. Two threshold values were tried: 10.0 and 2.0. The number of iterations in the estimation algorithm was 3 as the estimated frequencies seem to converge rather quickly.

In AUTOMATICALLY DISCARDED only forms of open-class lexemes (nouns, verbs, adjectives, adverbs and exclamations) were discarded as preliminary experiments had indicated that it does not pay to discard other types of lexemes.

AUTOMATICALLY DISCARDED was only slightly better than baseline ALL FORMS. The advantage is, however, more significant when the translations are judged subjectively as in AUTOMATICALLY DISCARDED many bizarre equivalents are avoided. One example is given below.

INPUT SENTENCE *Obecny kryzys przyczyni się do ograniczenia kredytów i już wywindował ceny, bez względu na to, jaką decyzję podejmie jutro Bank Centralny.*

REFERENCE TRANSLATION *The current crisis will have the effect of further restricting credit and it has already started to push prices up in anticipation, irrespective of the decision the Central Bank takes tomorrow.*

ALL FORMS *The present crisis will contribute to a credit squeeze and already hauled up prices, common lilac of the account of it, what decision the central bank will take the tomorrow.*

AUTOMATICALLY DISCARDED ($\theta = 10.0$) *The present crisis will contribute to a credit squeeze and already jacked prices up, irrespective of it, what decision tomorrow the central bank will make.*

Furthermore, AUTOMATICALLY DISCARDED is significantly faster than ALL FORMS as unwanted interpretations are avoided in parsing.

	Meteor	time	discarded forms
ALL FORMS	0.3081	402m19s	0
MANUALLY DISCARDED	0.3106	339m16s	124532
AUTOMATICALLY DISCARDED ($\theta = 10.0$)	0.3100	343m29s	27449
AUTOMATICALLY DISCARDED ($\theta = 2.0$)	0.3094	275m20s	54874

Table 2: Results of machine translation evaluation.

5. Anomaly detection

The quantitative data gathered while estimating the frequencies of inflected forms can be used to trace anomalies in the lexicon. Namely, form ϕ can be suspected of being anomalous if $f(\phi)/f(l(\phi))$ deviates significantly from $\delta(l(\phi), m(\phi))$. Some examples of anomalies found in the lexicon of Translatica system in this way:

- popular real-word spell checking errors overlapping with rare forms of infrequent lexemes, e.g. *naszczycie*, *wmieście*, *ujecie*, *pojecie*, *rzeczenie*,
- missing surnames (e.g. *Stępień*, *Smoleń*) or names of cities (e.g. *Augustynów*, *Sokolniki*) – such gaps in the lexicon had not been detected because the names were homonymous with rare forms already present in the lexicon,
- very frequent foreign words homonymous with rare Polish forms, e.g. *der* is theoretically the genitive plural of *dera* (= *rug*), but it is much more probable for it to be a fragment of a German or Dutch name used in a Polish text (e.g. *Der Spiegel*),
- various errors in the lexicon, e.g. *kranik* (= *little tap*) tagged incorrectly as a feminine noun, missing suppletive forms *latka* for *roczek* (= *diminutive year*), missing forms *burmistrzowie*, *mężowie*.

6. Conclusions and Future Work

We showed that it is possible to estimate the frequencies of inflected forms without a corpus using only a raw frequency list of words at least as far as discarding spurious homonyms is concerned.

It should be emphasised that not all frequencies could be correctly estimated using the algorithm described in Section : if two or more forms are always expressed with the same word in the given lexeme class then it is not possible to discriminate the estimations for such forms, e.g. the vocative singular for neuter nouns is always the same as the nominative singular and the accusative singular, hence the estimated frequencies for these 3 forms will always be the same, which is obviously incorrect as the frequency of vocatives is much lower than the frequency of nominatives and accusatives. The solution would be to consider "superclasses" of lexemes, e.g. the class of all nouns instead of the class of neuter nouns – the vocatives for feminine and masculine may differ from nominatives and accusatives and the lower frequency of vocatives can become apparent in this way.

7. Acknowledgment

The paper is based on research funded by the Polish Ministry of Science and Higher Education (Grant No N N516 480540).

References

Denkowski, M. and Lavie A. (2011): Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Graliński, F. (2000): Hasłowanie korpusu polskich tekstów informatycznych (1,2 mln słów) – raport. *Speech and Language Technology. Technologia Mowy i Języka*, t. 4, cz. II:147–153.

Jassem, K. (2006): *Przetwarzanie tekstów polskich w systemie tłumaczenia automatycznego POLENG*. Poznań: Wydawnictwo Naukowe UAM.

Jassem, K., Lison M., Graliński F., and Rutkowski B. (1998): A Polish-to-English electronic dictionary designed for the purposes of MT. In Frank van Eynde (ed.), *10th European Summer School in Logic, Languages and Information. Workshop on Machine Translation*. Saarbrücken: Universität Saarlandes.

Koehn, P. (2005): Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

Świdziński, M., Derwojedowa M., and Rudolf M. (2003): Dehomonimizacja i desynkretyzacja w procesie automatycznego przetwarzania wielkich korpusów tekstów polskich. *Biuletyn Polskiego Towarzystwa Językoznawczego*, LVIII:187–199.