Towards the Adequate Evaluation of Morphosyntactic Taggers

Szymon Acedański
Institute of Computer Science,
Polish Academy of Sciences
Institute of Informatics,
University of Warsaw
accek@mimuw.edu.pl

Adam Przepiórkowski
Institute of Computer Science,
Polish Academy of Sciences
Institute of Informatics,
University of Warsaw
adamp@ipipan.waw.pl

Abstract

There exists a well-established and almost unanimously adopted measure of tagger performance, namely, accuracy. Although it is perfectly adequate for small tagsets and typical approaches to disambiguation, we show that it is deficient when applied to rich morphological tagsets and propose various extensions designed to better correlate with the real usefulness of the tagger.

1 Introduction

Part-of-Speech (PoS) tagging is probably the most common and best researched NLP task, the first step in many higher level processing solutions such as parsing, but also information retrieval, speech recognition and machine translation. There are also well established evaluation measures, the foremost of which is accuracy, i.e., the percent of words for which the tagger assigns the correct — in the sense of some gold standard — interpretation.

Accuracy works well for the original PoS tagging task, where each word is assumed to have exactly one correct tag, and where the information carried by a tag is limited roughly to the PoS of the word and only very little morphosyntactic information, as in typical tagsets for English. However, there are two cases where accuracy becomes less than adequate: the situation where the gold standard and / or the tagging results contain multiple tags marked as correct for a single word, and the use of a rich morphosyntactic (or morphological) tagset.

The first possibility is discussed in detail in (Karwańska and Przepiórkowski, 2009), but the need for an evaluation measure for taggers which do not necessarily fully disambiguate PoS was already noted in (van Halteren, 1999), where the use of standard information retrieval measures precision and recall (as well as their harmonic mean, the F-measure) is proposed. Other natural generalisations of the accuracy measure, able to deal with non-unique tags either in the gold standard[1] or in the tagging results, are proposed in (Karwańska and Przepiórkowski, 2009).

Standard accuracy is less than adequate also in case of rich morphosyntactic tagsets, where the full tag carries information not only about PoS, but also about case, number, gender, etc. Such tagsets are common for Slavic languages, but also for Hungarian, Arabic and other languages. For example, according to one commonly used Polish tagset (Przepiórkowski and Woliński, 2003), the form *uda* has the following interpretations: `fin:sg:ter:perf` (a finite singular 3rd person perfective form of the verb UDAĆ 'pretend'), `subst:pl:nom:n` and

[1] There are cases were it makes sense to manually assign a number of tags as correct to a given word, as any decision would be fully arbitrary, regardless of the amount of context and world knowledge available. For example, in some Slavic languages, incl. Polish, there are verbs which optionally subcategorise for an accusative or a genitive complement, without any variation in meaning, and there are nouns which are syncretic between these two cases, so for such "verb + $\text{noun}_{\text{acc/gen}}$" sequences it is impossible to fully disambiguate case; see also (Oliva, 2001).

subst:pl:acc:n (nominative or accusative plural form of the neuter noun UDO 'thigh'). Now, assuming that the right interpretation in a given context is subst:pl:acc:n, accuracy will equally harshly penalise the other nominal interpretation (subst:pl:nom:n), which shares with the correct interpretation not only PoS, but also the values of gender and number, and the completely irrelevant verbal interpretation. A more accurate tagger evaluation measure should distinguish these two non-optimal assignments and treat subst:pl:nom:n as partially correct.

Similarly, the Polish tagset mentioned above distinguishes between nouns and gerunds, with some forms actually ambiguous between these two interpretations. For example, *zadanie* may be interpreted as a nominative or accusative form of the noun ZADANIE 'task', or a nominative or accusative form of the gerund derived from the verb ZADAĆ 'assign'. Since gerunds and nouns have very similar distributions, any error in the assignment of part of speech, noun vs. gerund, will most probably not matter for a parser of Polish — it will still be able to construct the right tree, provided the case is correctly disambiguated. However, the "all-or-nothing" nature of the accuracy measure regards the tag differing from the correct one only in part of speech or in case as harshly, as it would regard an utterly wrong interpretation, say, as an adverb.

In what follows we propose various evaluation measures which differentiate between better and worse incorrect interpretations, cf. § 2. The implementation of two such measures is described in § 3. Finally, § 4 concludes the paper.

2 Proposed Measures

2.1 Full Interpretations and PoS

The first step towards a better accuracy measure might consist in calculating two accuracy measures: one for full tags, and the other only for fragments of tags representing parts of speech. Two taggers wrongly assigning either fin:sg:ter:perf (T1) or subst:pl:nom:n (T2) instead of the correct subst:pl:acc:n would fare equally well with respect to the tag-level accuracy, but T2 would be — rightly — evaluated as better with respect to the PoS-level accuracy.

The second example given in § 1 shows, however, that the problem is more general and that a tagger which gets the PoS wrong (say, gerund instead of noun) but all the relevant categories (case, number, gender) right may actually be more useful in practice than the one that gets the PoS right at the cost of confusing cases (say, accusative instead of nominative).

2.2 Positional Accuracy

A generalisation of the idea of looking separately at parts of speech is to split tags into their components (or positions) and measure the correctness of the tag by calculating the F-measure. For example, if the (perfective, affirmative) gerundial interpretation ger:sg:nom:n:perf:aff is assigned instead of the correct nominal interpretation subst:sg:nom:n, the tags agree on 3 positions (sg, nom, n), so the precision is $\frac{3}{6}$, the recall — $\frac{3}{4}$, which gives the F-measure of 0.6. Obviously, the assignment of the correct interpretation results in F-measure equal 1.0, and the completely wrong interpretation gives F-measure 0.0. Taking these values instead of the "all-or-nothing" 0 or 1, accuracy is reinterpreted as the average F-measure over all tag assignments.

Note that while this measure, let us call it *positional accuracy* (PA), is more fine-grained than the standard accuracy, it wrongly treats all components of tags as of equal importance and difficulty. For example, there are many case syncretisms in Polish, but practically no ambiguities concerning the category of negation (see the value aff above), so case is inherently much more difficult than negation, and also much more important for syntactic parsing, and as such it should carry more weight when evaluating tagging results.

2.3 Weighted Positional Accuracy

In the current section we make a simplifying assumption that weights of positions are absolute, rather than conditional, i.e., that the weight of, say, case does not depend on part of speech, word or context. Once the weights are attained, weighted precision and recall may be used as in the following example.

Assume that PoS, case, number and gender have the same weight, say 2.0, which is 4 times larger than that of any other category. Then, in case `ger:sg:nom:n:perf:aff` is assigned instead of the correct `subst:sg:nom:n`, precision and recall are given by:

$$P = \frac{3 \times 2.0}{4 \times 2.0 + 2 \times 0.5} = \frac{2}{3},$$
$$R = \frac{3 \times 2.0}{4 \times 2.0} = \frac{3}{4}.$$

This results in a higher F-measure than in case of non-weighted positional accuracy.

The following subsections propose various ways in which the importance of particular grammatical categories and of the part of speech may be estimated.

2.3.1 Average Ambiguity

The average number of morphosyntactic interpretations per word is sometimes given as a rough measure of the difficulty of tagging. For example, tagging English texts with the Penn Treebank tagset is easier than tagging Czech or Polish, as the average number of possible tags per word is 2.32 in English (Hajič, 2004, p. 171), while it is 3.65 (Hajič and Hladká, 1997, p. 113) and 3.32 (Przepiórkowski, 2008, p. 44) for common tagsets for Czech and Polish, respectively.

By analogy, one measure of the difficulty of assigning the right value of a given category or part of speech is the average number of different values of the category per word.

2.3.2 Importance for Parsing

All measures mentioned so far are *intrinsic* (*in vitro*) evaluation measures, independent — but hopefully correlated with — the usefulness of the results in particular applications. On the other hand, *extrinsic* (*in vivo*) evaluation estimates the usefulness of tagging in larger systems, e.g., in parsers. Full-scale extrinsic evaluation is rarely used, as it is much more costly and often requires user evaluation of the end system.

In this and the next subsections we propose evaluation measures which combine the advantages of both approaches. They are variants of the weighted positional accuracy (WPA) measure, where weights correspond to the usefulness of a given category (or PoS) for a particular task.

Probably the most common task taking advantage of morphosyntactic tagging is syntactic parsing. Here, weights should indicate to what extent the parser relies on PoS and particular categories to arrive at the correct parse. Such weights may be estimated from an automatically parsed corpus in the following way:

for each category (including PoS) c **do**
  $\mathrm{count}(c) = 0$ {Initialise counts.}
end for
for each sentence s **do**
  for each rule r used in s **do**
    for each terminal symbol (word) t in the RHS of r **do**
      for each category c referred to by r in t **do**
        increase $\mathrm{count}(c)$
      end for
    end for
  end for
end for
{Use $\mathrm{count}(c)$'s as weights.}

In prose: whenever a syntactic rule is used, increase counts of all morphosyntactic categories (incl. PoS) mentioned in the terminal symbols occurring in this rule. These counts may be normalised or used directly as weights.

We assume here that either the parser produces a single parse for any sentence (assumption realistic only in case of shallow parsers), or that the best or at least most probable parse may be selected automatically, as in case of probabilistic grammars, or that parses are disambiguated manually. In case only a non-probabilistic deep parser is available, and parses are not disambiguated manually, the Expectation-Maximisation method may be used to select a probable parse (Dębowski, 2009) or all parses might be taken into account.

Note that, once a parser is available, such weights may be calculated automatically and used repeatedly for tagger evaluation, so the cost of using this measure is not significantly higher than the cost of intrinsic measures, while at the same time the correlation of the evaluation results with the extrinsic application is much higher.

2.3.3 Importance for Corpus Search

The final variant (many more are imaginable) of WPA that we would like to describe here concerns another application of tagging, namely, for the annotation of corpora. Various corpus search engines, including the IMS Open Corpus Workbench (`http://cwb.sourceforge.net/`) and Poliqarp (`http://poliqarp.sourceforge.net/`) allow the user to search for particular parts of speech and grammatical categories. Obviously, the tagger should maximise the quality of the disambiguation of those categories which occur frequently in corpus queries, i.e., the weights should correspond to the frequencies of particular categories (and PoS) in user queries. Note that the only resource needed to calculate weights are the logs of a corpus search engine.

An experiment involving an implementation of this measure is described in detail in § 3.

2.4 Conditional Weighted Positional Accuracy

The importance and difficulty of a category may depend on the part of speech. For example, after case syncretisms, gender ambiguity is one of the main problems for the current taggers of Polish. But this problem concerns mainly pronouns and adjectives, where the systematic gender syncretism is high. On the other hand, nouns do not inflect for gender, so only some nominal forms are ambiguous with respect to gender. Moreover, gerunds, which also bear gender, are uniformly neuter, so here part of speech alone uniquely determines the value of this category.

A straightforward extension of WPA capitalising on these observations is what we call *conditional weighted positional accuracy* (CWPA), where weights of morphosyntactic categories are conditioned on PoS.

Note that not all variants of WPA may be easily generalised to CWPA; although such an extension is obvious for the average ambiguity (§ 2.3.1), it is less clear for the other two variants. For parsing-related WPA, we assume that, even if a given rule does not mention the PoS of a terminal symbol,[2] that PoS may be read off the parse tree, so the conditional weights may still be calculated. On the other hand, logs of a corpus search engine are typically not sufficient to calculate such conditional weights; e.g., a query for a sequence of 5 genitive words occurring in logs would have to be rerun on the corpus again in order to find out parts of speech of the returned 5-word sequences. For a large number of queries on a large corpus, this is a potentially costly operation.

It is also not immediately clear how to generalise precision and recall from WPA to CWPA. Returning to the example above, where t_1 = `ger:sg:nom:n:perf:aff` is assigned instead of the correct t_2 = `subst:sg:nom:n`, we note that the weights of number, case and gender may now (and should, at least in case of gender!) be different for the two parts of speech involved. Hence, precision needs to be calculated with respect to the weights for the automatically assigned part of speech, and recall — taking into account weights for the gold standard part of speech:

$$P = \frac{\delta_{t_1^* t_2^*} w(t_1^*) + \sum_{c \in C(t_1,t_2)} \delta_{t_1^c t_2^c} w(c|t_1^*)}{w(t_1^*) + \sum_{c \in C(t_1)} w(c|t_1^*)},$$
$$R = \frac{\delta_{t_1^* t_2^*} w(t_2^*) + \sum_{c \in C(t_1,t_2)} \delta_{t_1^c t_2^c} w(c|t_2^*)}{w(t_2^*) + \sum_{c \in C(t_2)} w(c|t_2^*)},$$

where t^* is the PoS of tag t, $w(p)$ is the weight of the part of speech p, $w(c|p)$ is the conditional weight of the category c for PoS p, $C(t)$ is the set of morphosyntactic categories of tag t, $C(t_1, t_2)$ is the set of morphosyntactic categories common to tags t_1 and t_2, t^c is the value of category c in tag t, and δ_{ij} is the Kronecker delta (equal to 1 if $i = j$, and to 0 otherwise). Hence, for the example above, these formulas may be simplified to:

$$P = \frac{\sum_{c \in \{n,c,g\}} w(c|\texttt{ger})}{w(\texttt{ger}) + \sum_{c \in \{n,c,g,a,neg\}} w(c|\texttt{ger})},$$
$$R = \frac{\sum_{c \in \{n,c,g\}} w(c|\texttt{subst})}{w(\texttt{subst}) + \sum_{c \in \{n,c,g\}} w(c|\texttt{subst})},$$

where *n*, *c*, *g*, *a* and *neg* stand for number, case, gender, aspect and negation.

[2] For example, in unification grammars and constraint-based grammars a terminal may be identified only by the values of some of its categories, as in the following simple rule, specifying prepositional phrases as a preposition governing a specific case and a non-empty sequence of words bearing that case: $\text{PP}_{\text{case=C}} \rightarrow \text{P}_{\text{case=C}}\ \text{X}^{+}_{\text{case=C}}$.

3 Experiment

To evaluate behaviour of the proposed metrics, a number of experiments were performed using the manually disambiguated part of the IPI PAN Corpus of Polish (Przepiórkowski, 2005). This subcorpus consists of 880 000 segments. Two taggers of Polish were tested. TaKIPI (Piasecki and Godlewski, 2006) is a tagger which was used for automatic disambiguation of the remaining part of the aforementioned corpus. It is a statistical classifier based on decision trees combined with some automatically extracted, hand-crafted rules. This tagger by default sometimes assigns more than one tag to a segment, what is consistent with the golden standard. There is a setting which allows this behaviour to be switched off. This tagger was tested with both settings. The other tagger is a prototype version of this Brill tagger, presented by Acedański and Gołuchowski in (Acedański and Gołuchowski, 2009).

For comparison, four metrics were used: standard metrics for full tags and only parts of speech, as well as Positional Accuracy and Weighted Positional Accuracy. For the last measure, the weights were obtained by analysing logs of user queries of the Poliqarp corpus search engine. Occurrences of queries involving particular grammatical categories were counted and used as weights. Obtained results are presented in Table 1.

Table 1: Occurrences of particular grammatical categories in query logs of the Poliqarp corpus search engine.

Category	# occurrences
POS	37771
CASE	14055
NUMBER	2074
GENDER	552
ASPECT	222
PERSON	186
DEGREE	81
ACCOMMODABILITY	25
POST-PREP.	8
NEGATION	7
ACCENTABILITY	5
AGGLUTINATION	4

3.1 Scored information retrieval metrics

In § 2 a number of methods of assigning a score to a pair of tags were presented. From now on, let name them *scoring functions*. One could use them directly for evaluation, given that both the tagger and the golden standard always assign a single interpretation to each segment. This is not the case for the corpus we use, hence we propose generalisation of standard information retrieval metrics (precision, recall and F-measure) as well as strong and weak correctness (Karwańska and Przepiórkowski, 2009) to account for scoring functions.

Denote by T_i and G_i the sets of tags assigned by the tagger and the golden standard, accordingly, to the i-th segment of the tagged text. The set of all tags in the tagset is denoted by $\mathbf{T}$. The scoring function used is $score\colon \mathbf{T} \times \mathbf{T} \to [0, 1]$. Also, to save up on notation, we define

$$score(t, A) := \max_{t' \in A} score(t, t')$$

Now, given the text has n segments, we take

$$P = \frac{\sum_{i=1}^{n} \sum_{t \in T_i} score(t, G_i)}{\sum_{i=1}^{n} |T_i|}$$

$$R = \frac{\sum_{i=1}^{n} \sum_{g \in G_i} score(g, T_i)}{\sum_{i=1}^{n} |G_i|}$$

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

$$WC = \frac{\sum_{i=1}^{n} \max_{t \in T_i} score(t, G_i)}{n}$$

$$SC = \frac{\sum_{i=1}^{n} \min(\{score(t, G_i) \colon t \in T_i\} \cup \{score(g, T_i) \colon g \in G_i\})}{n}$$

Intuitions for scored precision and recall are that precision specifies the percent of tags assigned by the tagger which have a high score with some corresponding golden tag. Analogously recall estimates the percent of golden tags which have high scores with some corresponding tag assigned by the tagger. The definition of recall is slightly different than proposed by Ziółko et al. (Ziółko et al., 2007) so that recall is never greater than one.[3]

[3] For example if the golden standard specifies a single tag and the tagger determines two tags which all score 0.6 when compared with the golden, then if we used equations from Ziółko et al., we would get the recall of 1.2.

3.2 Evaluation results

Now the taggers were trained on the same data consisting of 90% segments of the corpus and then tested on the remaining 10%. Results were 10-fold cross-validated. They are presented in Tables 2, 3, 4 and 5.

As expected, the values obtained with PA and WPA fall between the numbers for standard metrics calculated with full tags and only the part of speech. What is worth observing is that the use of WPA makes values for scored precision and recall much closer together. This can be justified by the fact that the golden standard relatively frequently contains more than one interpretation for some tags, which differ only in values of less important grammatical categories. WPA is resilient to such situations.

One may argue that such scoring functions may hide a large number of tagging mistakes occurring in low-weighted categories. But this is not the case as the clearly most common tagging errors reported in both (Piasecki and Godlewski, 2006) and (Acedański and Gołuchowski, 2009) are for CASE, GENDER and NUMBER. Also, the motivation for weighting grammatical categories is to actually ignore errors in not important ones. To be fair, though, one should make sure that the weights used for evaluation match the actual application domain of the analysed tagger, and if no specific domain is known, using a number of measures is recommended.

It should also be noted that for classic information retrieval metrics, the result of weak correctness for TaKIPI is more similar to 92.55% reported by the authors (Piasecki and Godlewski, 2006) than 91.30% shown in (Karwańska and Przepiórkowski, 2009) despite using the same test corpus and very similar methodology[4] as the second paper presents.

4 Conclusion

This paper stems from the observation that the commonly used measure for tagger evaluation, i.e., accuracy, does not distinguish between completely incorrect and partially correct interpretations, even though the latter may be sufficient for some applications. We proposed a way of grading tag assignments, by weighting the importance of particular categories (case, number, etc.) and the part of speech. Three variants of the weighted positional accuracy were presented: one intrinsic and two application-oriented, and an extension of WPA to conditional WPA was discussed. The variant of WPA related to the needs of the users of a corpus search engine for the National Corpus of Polish was implemented and its usefulness was demonstrated. We plan to implement the parsing-oriented WPA in the future.

We conclude that tagger evaluation is far from being a closed chapter and the time has come to adopt more subtle approaches than sheer accuracy, approaches able to cope with morphological richness and oriented towards real applications.

[4]The only difference was not contracting the grammatical category of ACCOMMODABILITY present for masculine numerals in the golden standard.

References

Acedański, Szymon and Konrad Gołuchowski. 2009. A morphosyntactic rule-based Brill tagger for Polish. In Kłopotek, Mieczysław A., Adam Przepiórkowski, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Advances in Intelligent Information Systems — Design and Applications*, pages 67–76. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Dębowski, Łukasz. 2009. Valence extraction using the EM selection and co-occurrence matrices. *Language Resources and Evaluation*, 43:301–327.

Hajič, Jan and Barbora Hladká. 1997. Probabilistic and rule-based tagger of an inflective language - a comparison. In *Proceedings of the 5th Applied Natural Language Processing Conference*, pages 111–118, Washington, DC. ACL.

Hajič, Jan. 2004. *Disambiguation of Rich Inflection*. Karolinum Press, Prague.

Janus, Daniel and Adam Przepiórkowski. 2007. Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 85–88, Prague.

Karwańska, Danuta and Adam Przepiórkowski. 2009. On the evaluation of two Polish taggers. In Goźdź-Roszkowski, Stanisław, editor, *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main. Peter Lang. Forthcoming.

Table 2: Evaluation results — standard information retrieval metrics, full tags

Tagger	C (%)	WC (%)	P (%)	R (%)	F (%)
TaKIPI — defaults	87.67%	92.10%	89.93%	84.72%	87.25%
TaKIPI — one tag per seg.	88.68%	91.06%	90.94%	83.78%	87.21%
Brill	90.01%	92.44%	92.26%	85.00%	88.49%

Table 3: Evaluation results — standard information retrieval metrics, PoS only

Tagger	C (%)	WC (%)	P (%)	R (%)	F (%)
TaKIPI — defaults	95.56%	97.52%	95.71%	97.61%	96.65%
TaKIPI — one tag per seg.	96.53%	96.54%	96.58%	96.71%	96.65%
Brill	98.17%	98.18%	98.20%	98.26%	98.23%

Table 4: Evaluation results — scored metrics, Positional Accuracy

Tagger	C (%)	WC (%)	P (%)	R (%)	F (%)
TaKIPI — defaults	95.23%	96.58%	95.69%	95.44%	95.57%
TaKIPI — one tag per seg.	95.69%	96.10%	96.12%	95.00%	95.56%
Brill	97.02%	97.43%	97.42%	96.27%	96.84%

Table 5: Evaluation results — scored metrics, Weighted PA, Poliqarp weights

Tagger	C (%)	WC (%)	P (%)	R (%)	F (%)
TaKIPI — defaults	95.20%	96.62%	95.34%	96.56%	95.95%
TaKIPI — one tag per seg.	95.88%	95.93%	95.97%	95.94%	95.95%
Brill	97.34%	97.40%	97.41%	97.34%	97.38%

Oliva, Karel. 2001. On retaining ambiguity in disambiguated corpora: Programmatic reflections on why's and how's. *TAL (Traitement Automatique des Langues)*, 42(2):487–500.

Piasecki, Maciej and Grzegorz Godlewski. 2006. Effective Architecture of the Polish Tagger. In Sojka, Petr, Ivan Kopecek, and Karel Pala, editors, *TSD*, volume 4188 of *Lecture Notes in Computer Science*, pages 213–220. Springer.

Przepiórkowski, Adam and Marcin Woliński. 2003. The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, pages 109–116.

Przepiórkowski, Adam. 2005. The IPI PAN Corpus in Numbers. In *Proceedings of the* 2nd Language & Technology Conference, Poznań, Poland.

Przepiórkowski, Adam. 2008. *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

van Halteren, Hans. 1999. Performance of taggers. In van Halteren, Hans, editor, *Syntactic Wordclass Tagging*, volume 9 of *Text, Speech and Language Technology*, pages 81–94. Kluwer, Dordrecht.

Ziółko, Bartosz, Suresh Manandhar, and Richard Wilson. 2007. Fuzzy Recall and Precision for Speech Segmentation Evaluation. In *Proceedings of 3rd Language & Technology Conference, Poznan, Poland*, October.