Icelandic Data Driven Part of Speech Tagging

Mark Dredze
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
mdredze@cis.upenn.edu

Joel Wallenberg
Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104
joelcw@babel.ling.upenn.edu

Abstract

Data driven POS tagging has achieved good performance for English, but can still lag behind linguistic rule based taggers for morphologically complex languages, such as Icelandic. We extend a statistical tagger to handle fine grained tagsets and improve over the best Icelandic POS tagger. Additionally, we develop a case tagger for non-local case and gender decisions. An error analysis of our system suggests future directions.

1 Introduction

While part of speech (POS) tagging for English is very accurate, languages with richer morphology demand complex tagsets that pose problems for data driven taggers. In this work we consider Icelandic, a language for which a linguistic rule-based method is the current state of the art, indicating the difficulty this language poses to learning systems. Like Arabic and Czech, other morphologically complex languages with large tagsets, Icelandic can overwhelm a statistical tagger with ambiguity and data sparsity.

Shen et al. (2007) presented a new framework for bidirectional sequence classification that achieved the best POS score for English. In this work, we evaluate their tagger on Icelandic and improve results with extensions for fine grained annotations. Additionally, we show that good performance can be achieved using a strictly data-driven learning approach without external linguistic resources (morphological analyzer, lexicons, etc.). Our system achieves the best performance to date on Icelandic, with insights that may help improve other morphologically rich languages.

After some related work, we describe Icelandic morphology followed by a review of previous approaches. We then apply a bidirectional tagger and extend it for fine grained languages. A tagger for case further improves results. We conclude with an analysis of remaining errors and challenges.

2 Related Work

Previous approaches to tagging morphologically complex languages with fine grained tagsets have considered Czech and Arabic. Khoja (2001) first introduced a tagger for Arabic, which has 131 tags, but subsequent work has collapsed the tagset to simplify tagging (Diab et al., 2004). Like previous Icelandic work (Loftsson, 2007), morphological analyzers disambiguate words before statistical tagging in Arabic (Habash and Rambow, 2005) and Czech (Hajič and Hladká, 1998). This general approach has led to the serial combination of rule based and statistical taggers for efficiency and accuracy (Hajič et al., 2001). While our tagger could be combined with these linguistic resources as well, as in Loftsson (2007), we show state of the art performance without these resources. Another approach to fine-grained tagging captures grammatical structures with tree-based tags, such as "supertags" in the tree-adjoining grammar of Bangalore and Joshi (1999).

3 Icelandic Morphology

Icelandic is notable for its morphological richness. Verbs potentially show as many as 54 different forms depending on tense, mood, voice, person and

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 33–36,
Columbus, Ohio, USA, June 2008.

number. A highly productive class of verbs also show stem vowel alternations reminiscent of Semitic verb morphology (Arabic). Noun morphology exhibits a robust case system; nouns may appear in as many as 16 different forms. The four-case system of Icelandic is similar to that of the Slavic languages (Czech), with case morphology also appearing on elements which agree in case with nouns. However, unlike Czech, case frequently does not convey distinct meaning in Icelandic as it is often determined by elements such as the governing verb in a clause (non-local information). Therefore, while Icelandic case looks formally like Slavic and presents similar challenges for POS tagging, it also may be syntactically-determined, as in Standard Arabic. Icelandic word-order allows a very limited form of scrambling, but does not produce the variety of permutations allowed in Slavic languages. This combination of morphological complexity and syntactic constraint makes Icelandic a good case study for statistical POS tagging techniques.

The morphology necessitates the large extended tagset developed for the Icelandic Frequency Dictionary (Íslensk orðtíðnibók/IFD), a corpus of roughly 590,000 tokens (Pind et al., 1991). We use the 10 IFD data splits produced by Helgadóttir (2004), where the first nine splits are used for evaluation and the tenth for model development. Tags are comprised of up to six elements, such as word class, gender, number, and case, yielding a total of 639 tags, not all of which occur in the training data.

4 Previous Approaches

Helgadóttir (2004) evaluated several data-driven models for Icelandic, including MXPost, a maximum entropy tagger, and TnT, a trigram HMM; both did considerably worse than on English. Icelandic poses significant challenges: data sparseness, non-local tag dependencies, and 136,264 observed trigram sequences make discriminative sequence models, such as CRFs, prohibitively expensive. Given these challenges, the most successful tagger is IceTagger (Loftsson, 2007), a linguistic rule based system with several linguistic resources: a morphological analyzer, a series of local rules and heuristics for handling PPs, verbs, and forcing agreement. Loftsson also improves TnT by integrating a morphological analyzer (TnT*).

Despite these challenges, data driven taggers have several advantages. Learning systems can be easily applied to new corpora, tagsets, or languages and can accommodate integration of other systems (including rule based) or new linguistic resources, such as those used by Loftsson. Therefore, we seek a learning system that can handle these challenges.

5 Bidirectional Sequence Classification

Bidirectional POS tagging (Shen et al., 2007), the current state of the art for English, has some properties that make it appropriate for Icelandic. For example, it can be trained quickly with online learning and does not use tag trigrams, which reduces data sparsity and the cost of learning. It can also allow long range dependencies, which we consider below.

Bidirectional classification uses a perceptron style classifier to assign potential POS tags (hypotheses) to each word using standard POS features and some additional local context features. On each round, the algorithm selects the highest scoring hypothesis and assigns the guessed tag. Unassigned words in the context are reevaluated with this new information. If an incorrect hypothesis is selected during training, the algorithm promotes the score of the correct hypothesis and demotes the selected one. See Shen *et al.* for a detailed explanation.

We begin with a direct application of the bidirectional tagger to Icelandic using a beam of one and the same parameters and features as Shen *et al.* On the development split the tagger achieved an accuracy of 91.61%, which is competitive with the best Icelandic systems. However, test evaluation is not possible due to the prohibitive cost of training the tagger on nine splits; training took almost 4 days on an AMD Opteron 2.8 GHz machine.

Tagset size poses a problem since the tagger must evaluate over 600 options to select the top tag for a word. The tagger rescores the local context after a tag is committed or all untagged words if the classifier is updated. This also highlights a problem with the learning model itself. The tagger uses a one vs. all multi-class strategy, requiring a correct tag to have higher score than every other tag to be selected. While this is plausible for a small number of labels, it overly constrains an Icelandic tagger.

	Accuracy			Train
Tagger	*All*	*Known*	*Unkn.*	*Time*
Bidir	91.61	93.21	69.76	90:27
Bidir+WC	91.98	93.58	70.10	12:20
Bidir+WC+CT	92.36	93.93	70.95	14:02

Table 1: Results on development data. Accuracy is measured by exact match with the gold tag. About 7% of tokens are unknown at test time.

As with most languages, it is relatively simple to assign word class (noun, verb, etc.) and we use this property to divide the tagset into separate learning problems. First, the tagger classifies a word according to one of the eleven word classes. Next, it selects and evaluates all tags consistent with that class. When an incorrect selection is updated, the word class classifier is updated only if it was mistaken as well. The result is a dramatic reduction in the number of tags considered at each step. For some languages, it may make sense to consider further reductions, but not for Icelandic since case, gender, and number decisions are interdependent. Additionally, by learning word class and tag separately, a correct tag need only score higher than other tags of the same word class, not all 639. Furthermore, collapsing tags into word class groups increases training data, allowing the model to generalize features over all tags in a class instead of learning each tag separately (a form of parameter tying).

Training time dropped to 12 hours with the bidirectional word class (WC) tagger and learning performance increased to 91.98% (table 1). Word class accuracy, already quite high at 97.98%, increased to 98.34%, indicating that the tagger can quickly filter out most inappropriate tags. The reduced training cost allowed for test data evaluation, yielding 91.68%, which is a 12.97% relative reduction in error over the best pure data driven model (TnT) and a 1.65% reduction over the best model (IceTagger).

6 Case Tagger

Examining tagger error reveals that most mistakes are caused by case confusion on nouns (84.61% accuracy), adjectives (76.03%), and pronouns (90.67%); these account for 40% of the corpus. While there are 16 case-number-definiteness combinations in the noun morphology, a noun might realize several combinations with a single phonological/orthographic form (case-syncretism). Mistakes in noun case lead to further mistakes for categories which agree with nouns, e.g. adjectives. Assigning appropriate case for nouns is important for a number of other tagging decisions, but often the noun's case provides little or no information about the identity of other tags. It is in this situation that the tagger makes most case-assignment errors. Therefore, while accuracy depends on correct case assignment for these nouns, other tags are mostly unaffected.

One approach to correcting these errors is to introduce long range dependencies, such as those used by IceTagger. While normally hard to add to a learning system, bidirectional learning provides a natural framework since non-local features can be added once a tag has been committed. To allow dependencies on all other tag assignments, and because correcting the remaining case assignments is unlikely to improve other tags, we constructed a separate bidirectional case tagger (CT) that retags case on nouns, adjectives and pronouns. [1] Since gender is important as it relates to case, it is retagged as well. The CT takes a fully tagged sentence from the POS tagger and retags case and gender to nouns, adjectives and pronouns. The CT uses the same features as the POS tagger, but it now has access to all predicted tags. Additionally, we develop several non-local features.

Many case decisions are entirely idiosyncratic, even from the point of view of human language-learners. Some simple transitive verbs in Icelandic arbitrarily require their objects to appear in dative or genitive case, rather than the usual accusative. This arbitrary case-assignment adds no additional meaning, and this set of idiosyncratic verbs is memorized by speakers. A statistical tagger likewise must memorize these verbs based on examples in the training data. To aid generalization, verb-forms were augmented by verb-stems features as described in Dredze and Wallenberg (2008): e.g., the verb forms `dveldi, dvaldi, dvelst, dvelur` all mapped to the stem `dv*l` (*dvelja* "dwell"). The tagger used non-local features, such as the preceding verb's (predicted) tag, gender, case, stem, and nouns within the clause boundary as indicated by

[1] We considered adding case tagging features to and removing case decisions from the tagger; both hurt performance.

Tagger	*All*	*Known*	*Unknown*
MXPost	89.08	91.04	62.50
TnT	90.44	91.82	71.68
TnT*	91.18	92.53	72.75
IceTagger	91.54	92.74	75.09
Bidir+WC	91.68	93.32	69.25
Bidir+WC+CT	**92.06**	93.70	69.74

Table 2: Results on test data.

the tags `cn` (complementizer) or `ct` (relativizer) (Dredze and Wallenberg, 2008).

The CT was used to correct the output of the tagger after training on the corresponding train split. The CT improved results yielding a new best accuracy of 92.06%, a 16.95% and 12.53% reduction over the best data driven and rule systems.

7 Remaining Challenges

We have shown that a data driven approach can achieve state of the art performance on highly inflected languages by extending bidirectional learning to fine grained tagsets and designing a bidirectional non-local case tagger. We conclude with an error analysis to provide future direction.

The tagger is particularly weak on unknown words, a problem caused by case-syncretism and idiosyncratic case-assignment. Data driven taggers can only learn which verbs assign special object cases by observation in the training data. Some verbs and prepositions also assign case based on the meaning of the whole phrase. These are both serious challenges for data-driven methods and could be addressed with the integration of linguistic resources.

However, there is more work to be done on data driven methods. Mistakes in case-assignment due to case syncretism, especially in conjunction with idiosyncratic-case-assigning verbs, account for a large proportion of remaining errors. Verbs that take dative rather than accusative objects are a particular problem, such as mistaking accusative for dative feminine objects (10.6% of occurrences) or dative for accusative feminine objects (11.9%). A possible learning solution lies in combining POS tagging with syntactic parsing, allowing for the identification of clause boundaries, which may help disambiguate noun cases by deducing their grammatical function from that of other clausal constituents.

Additionally, idiosyncratic case-assignment could be learned from *unlabeled* data by finding unambiguous dative objects to identify idiosyncratic verbs. Furthermore, our tagger learns which prepositions idiosyncratically assign a single odd case (e.g. genitive) since prepositions are a smaller class and appear frequently in the corpus. This indicates that further work on data driven methods may still improve the state of the art.

8 Acknowledgments

We thank Hrafn Loftsson for sharing IceTagger and the datasplits, Libin Shen for his tagger, and the Árni Magnússon Institute for Icelandic Studies for access to the corpus.

References

Srinivas Bangalore and Arivand K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2).

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of arabic text: From raw text to base phrase chunks. In *NAACL/HLT*.

Mark Dredze and Joel Wallenberg. 2008. Further results and analysis of icelandic part of speech tagging. Technical Report MS-CIS-08-13, CIS Dept, University of Pennsylvania.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *ACL*.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: prediction of morphological categories for a rich, structured tagset. In *COLING*.

Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva, and Vladimír Petkevič. 2001. Serial combination of rules and statistics: a case study in czech tagging. In *ACL*.

Sigrun Helgadóttir. 2004. Testing data-driven learning algorithms for pos tagging of icelandic.

Shereen Khoja. 2001. Apt: Arabic part-of-speech tagger. In *NAACL Student Workshop*.

Hrafn Loftsson. 2007. Tagging icelandic text using a linguistic and a statistical tagger. In *NAACL/HLT*.

J Pind, F Magnússon, and S Briem. 1991. The icelandic frequency dictionary. Technical report, The Institute of Lexicography, University of Iceland.

Libin Shen, Giorgio Satta, and Aravind K. Joshi. 2007. Guided learning for bidirectional sequence classification. In *ACL*.