Tagging Icelandic text: A linguistic rule-based approach

Paper by Hrafn Loftsson

Presented by Haukur Kristinsson

What is the paper about?

Describes the design of a linguistic rule-based system for POS (Part of Speech) tagging Icelandic text

POS Tagging

- Labelling words with the appropriate
  - Word class
  - Morphological features
- Each label is called a tag and is from a tagset
- Program that performs the tagging is called a tagger
- Tagging text is needed for several NLP tasks
  - Grammar correction
  - Syntatactic parsing
  - Information extraction
  - Question-answering
  - Corpus annotation

Icelandic tag-set

- Main tagset, created during the making of the IFD 'Icelandic Frequency Dictionary'
  - Large tag-set (about 660 tags)
- First character denotes the word class (Noun, Adjective, Verb etc.)
- Additional characters (at most 5) describe morphological features
  - Gender (í. Kyn)
  - Number (í. Flrt/Eint)
  - Case (í. Fallbeyging)
  - Article And Proper Nouns (For Nouns) (í. Greinir/Heiti)
  - Declension and Degree (For Adjectives) (í. Beyging og stig lýsingaro.)
  - Mood – Person – Tense (For Verbs) (Í. Háttur – Persóna – Tíð)

Semantics of the tag-set

Semantics for nouns and adjectives

Char #	Category/ Feature	Symbol – semantics
1	Word class	**n**-noun, **l**-adjective
2	Gender	**k**-masculine, **v**-feminine, **h**-neuter, **x**-unspecified
3	Number	**e**-singular, **f**-plural
4	Case	**n**-nominative, **o**-accusative, **þ**-dative, **e**-genitive
5	Article	**g**-with suffixed definite article (nouns)
5	Declension	s-strong, **v**-weak (adjectives)
6	Proper noun	**m**-person name, **ö**-place name, s-other
6	Degree	**f**-positive, **m**-comparative, **e**-superlative (adjectives)

Semantics for verbs

Char #	Category/ Feature	Symbol – semantics
1	Word class	s-verb (except for past participle)
2	Mood	**n**-infinitive, **b**-imperative, **f**-indicative, **v**-subjunctive, s-supine, **l**-present participle
3	Voice	**g**-active, **m**-middle
4	Person	**1**-1st person, **2**-2nd person, **3**-3rd person
5	Number	**e**-singular, **f**-plural
6	Tense	**n**-present, **þ**-past

Example:

Untagged:
Fallegu hestarnir stukku

Tagged:
Fallegu/**lkfnvf**
hestarnir/**nkfng**
Stukku/**sfg3fg**

Function of a Tagger

- Remove ambiguity (lexical phase)
  - First, introduce the 'tag profile' for each word
    - Done by precompiled lexicon and a unknown word guesser
  - Second, do a morphical disambiguation on the word
- Two main methodologies to disambiguate
  - Data-driven
    - Uses pre-tagged training corpus
    - Machine learning to automaticlly derive a language model from the corpus
    - Less human effort
  - Linguistic rule-based approach (handcrafted)
    - Uses hand-crafted rules or constraints to eliminate appropriate POS tags (depending on the context)
    - More Human effor

Tagging methods

- In this research paper we discuss 2 methods
  - Data-driven tagging methods
    - Methods that are 'standard' today
    - Easier to develop
    - Taggers that we will be compared to IceTagger
  - Linguistic rule-based tagging methods
    - Methods that are used in IceTagger
    - Harder to develop
- Important to develop different approaches for a particular language
  - They produce uncorrelated errors
  - Can be used together with a simple voting to yield better results

Data-driven tagging methods

- Types of data-driven taggers used in this research
  - Probabilistic trigram taggers
    - Tag words by optimizing the product of lexical and contextual probabilities.
    - Trigram tagger based on Markov model (TnT Tagger)
    - Tagger based on maximum entropy approach (MXPOST Tagger)
  - Transformation-based learning approach tagger (fnTBL Tagger)
    - Rules based but not hand-crafted, rules acquired from a pre-tagged corpus

Linguistic rule-based tagging methods

- Purpose to tag a specific language
- Purpose of the rules
  - Assign tags to words depending on the context
  - Remove illegitimate tags from words based on context
- Time consuming task (because it can be many hand-crafted rules)

Unknown word guessing

- Main problem of a two-stage tagger
  - Guessing tag profile for unknown words.
- Constantly extending the lexicon to minimize unknown words not practical
  - New words constantly being introduced into a language
- Good quality unknown word guesser is essential to develop a high accuracy tagger.

Unknown word guessing

- Most unknown word guessers use
  - Morphological/Compound analysis
    - Analyzes morphologically related words already known to the lexicon
    - More accurate
  - Ending analysis
    - Analyzes solely on the word's ending
  - Combination of both

Tagging Icelandic

- Icelandic language is a morphologically complex language
  - Large tag-set
- Linguistic rule-based system for POS Icelandic text
- First we introduce the 'tag profile' for each word with
  - Pre-compiled lexicon
  - IceMorphy
- Main components
  - IceTagger, a disambiguator.
    - Uses about 175 rules along with heuristics
  - IceMorphy, the unknown word guesser.

IceMorphy

- Purpose to generate the tag profile for given word.
- It performs
  - Morphological analysis (Most accurate)
    - Classify the word as a member of morphological class
      - 18 morphological classes for nouns, 5 for adjectives and 5 for verbs
    - Class is guessed based on the words morphological suffix
      - After finding the suffix (and the word class) the stem is extracted from the word (stem+suffix)
      - All possible suffixes for the stem are generated and searched until finding a word in the same morphological class.
  - Compound analysis
    - Removes prefixes from the word and searches in the lexicon
      - If not it sends it to the morphological analysis.
    - Example: nýfæddur -> looks up 'fæddur' and gives 'nýfæddur' the same tag.

IceMorphy

- It Performs (continue..)
  - Ending analysis (Less accurate)
    - Used if nothing was found by morphological nor compound analysis fails
    - Uses the end of the word to look up in a ending lexicon (hand-written and generated ending from a corpus)
    - Example -> bleðillinn -> based on the ending 'llinn' we get the four tags 'nkeng_nkeog_lkensf_lkeosf' only the first tag is correct so you see how unaccurate it is
- Last important feature – Tagging profile gaps
  - When word has some missing tags in its set of possible tags.
  - For each noun, adjective or verb of a particular morphological class, IceMorphy generates all missing tags with all the methods above.
    - Konu 'woman' comes with only **nveo** tag, the methods detects from the suffix 'u' that it's a feminine noun class and it has the same form in singular accusative, dative and genitive. So it adds **nveþ** and **nvee** to the word

IceTagger – Disambiguation Process

- First step of the disambiguation is to **identify idioms** (í. Orðatiltæki)
  - F.ex. bigrams and trigrams (they often get tagged ambiguously)
    - For example: "of the", "in the", "to the" etc…
  - Identified by examining lexical forms of adjacent words
  - Extracted all trigrams from the IFD corpus that occurred at least ten times with the same tag sequence
  - Hand constructed a list of unambiguous bigrams from a test corpora based on IFD.
- Second step of the disambiguation is **identifying phrasal-verb**
  - Word that are adjacent in text (f.ex verb-particle pair: fara út 'go out')
    - Where the particle is an adverb (because it's associated with a particulate verb) but not a preposition
  - Automatically generated from IFD corpus

IceTagger – Disambiguation Process

- Third step is application of **local elimination rules**
  - Disambiguation based on a local context
  - Window of 5 words
    - Two words to the left and two words to the right
    - Focus word in the middle
    - L1/R1 L2/R2 denotes one and two to the left/right of the word
  - Purpose is to eliminate inappropriate tags from words
  - Example -> við vorum alltaf ein 'we were always alone'
    - við can have following five tags: **ao_aþ_fp1fn_aa_nkeo**
    - For example a rule **for preposition <condition> = R1.isOnlyWordClass(Verb)** eliminates prepositions tags in this context because the following word is a verb, leaving **fp1fn_aa_nkeo.**

IceTagger – Heuristics

- When disambiguation has finished every sentence is sent to the Heuristics module
- Its purpose is to perform
  - Grammatical function analysis
  - Guess prepositional phrases
  - Use acquired knowledge to force feature agreement where appropriate

IceTagger - Heuristics

- The Heuristics repeatedly scan each sentence and perform the following (in order)
  - 1. Mark prepositional phrases
  - 2. Mark verbs
  - 3. Mark subjects of verbs
  - 4. Force subject-verb agreement
  - 5. Mark objects of verbs
  - 6. Force subject-object agreement
  - 7. Force verb-object agreement
  - 8. Force agreement between nominal's
  - 9. Force prepositional phrase agreement

Heuristic Example

- Ég/**fp1en** fór/**sfg3eþ_sfg1eþ** svartar/**lvfosf_lvnsf** götur/**nvfo_nvfn** í/**aþ_ao** vesturátt/**nveo_nveþ**
- *1. Marks 'í vesturátt' as a prepositional phrase*
  - *'í' is an preposition and 'vesturátt' is a nominal.*
- *2. Marks 'fór' as an verb*
- *3. Marks 'ég' as a subject, as it is a subject of the verb fór.*
- *4. Removes **sfg3eþ** from 'fór'*
  - *the subject 'ég' is 1st person.*
- *5. Marks 'götur' as the object of the verb 'fór'*
- *7. Removes the nominative tag **nvfn** from 'götur'*
  - *The verb 'fór' demands an accustative (í. þf.) object (this is a rule obtained from a special lexicon that is made for helping verb-object aggreement)*
- *8. Removes nomitive (í. Nf.) tag **lvfnsf** from the adjective 'svartar'*
  - *The already disambiguated noun 'götur' (nominal) – Agreement between nominals.*

Heuristic Example

- Ég/**fp1en** fór/**sfg3eþ_sfg1eþ** svartar/**lvfosf_lvnsf** götur/**nvfo_nvfn** í/**aþ_ao** vesturátt/**nveo_nveþ**
  - 9. Removes the dative (í. Þgf.) tag **aþ** from preposition 'í' and the dative tag **nveþ** from the nominal 'vesturátt'.
    - The preposition pair fór-í governs accusative (í. Þf.) case
    - Rule obtained from a lexicon that is made specially to aid prepositional phrase agreement)
- Ég/**fp1en** fór/**sfg3eþ_~~sfg1eþ~~** svartar/**lvfosf_~~lvnsf~~** götur/**nvfo_~~nvfn~~** í/**~~aþ~~_ao** vesturátt/**nveo_~~nveþ~~**
- Ég/**fp1en** fór/**sfg3eþ** svartar/**lvfosf** götur/**nvfo** í/**ao** vesturátt/**nveo**

Evaluvation/Conclusion

- Compared Linguistic rule-based tagger (IceTagger) with IceMorphy to three state-of-the-art data-driven taggers
  - Obtained a higher accuracy when tagging Icelandic w. the large tagset
  - Main lexicon is obtained from the tagged corpus
  - The avarage tagging accuracy of IceTagger is 91.54%
  - The highest avarage tagging accuracy from the data-driven taggers is 90.44% (w. gap filling from IceMorphy 91.18%)
  - With combining IceTagger with 2 highest data-driven taggers (fnTBL and TnT) he accuracy raised to 92.95%.