Cross-lingual morphological analysis

Олег Сериков
Лоренцо Тони

Настя Хорошева
Влад Михайлов

Фран Тайерз

Похожие проекты

Initial Experiments in Data-Driven Morphological Analysis for Finnish (Silfverberg, Hulden)

программа осуществляет морфологический анализ для незнакомых финских слов

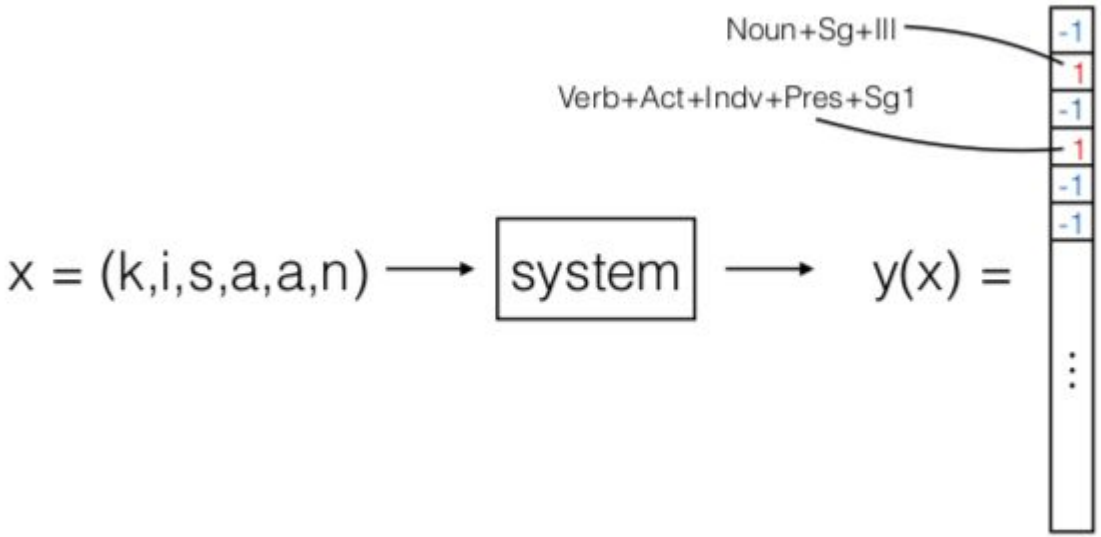
ввод: набор финских маркированных слов (UD Finnish Tree + OMorFi) + машинное обучение (долгая краткосрочная память)

слова = вектор со всеми возможными сетами морфологических признаков (у валидных y = 1, у остальных = -1)

F-Score All words 91.79 OOV words 49.89

Пример

kisaan (‘into the competition’ or ‘I am competing’)



Пример

Initial Experiments in Data-Driven Morphological Analysis for Finnish (Silfverberg, Hulden):

kisaan (‘into the competition’ or ‘I am competing’)

{Noun+Sg+Ill, Verb+Act+Indv+Pres+Sg1}

Цель

Modelling morphology of an unseen language with annotated data on related languages given.

Ожидаемый F-Score = 80%

(т.к. порой мы не можем использовать морфологические признаки, которые отсутствуют в уже имеющихся языках).

Материал: славянские языки

РИСКИ

Сложности:

- потенциальное отсутствие морфологических маркеров в train set
- разная морф. и типологическая структура языков внутри семьи
- разные алфавиты

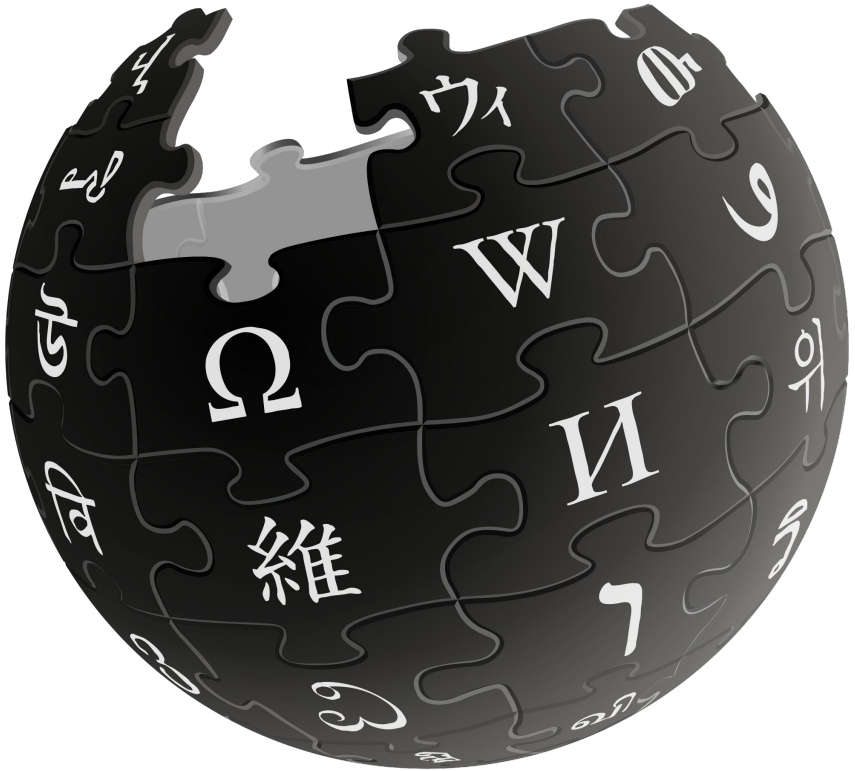
Потенциальные фейлы:

- upper-case letters
- OOV лексика
- омонимия

Материал проекта

Testing data - Wiki dumps:

- Русский
- Чешский
- Польский
- Украинский
- Беларусский
- Болгарский
- Македонский
- Словенский
- Сербский
- Хорватский
- Силезский



Материал проекта

- Apertium WikiExtractor
- Модули Apertium
- Топ 10 000 словоформ для каждого открытого лексического класса (<n>, <vblex>, <adj>)

```
1917781 ^года/год<n><m><nn><sg><gen>/год²<n><m><nn><sg><gen>/год²<n><m><nn><pl><acc>/год²<n><m><n
1460403 ^году/год<n><m><nn><sg><dat>/год<n><m><nn><sg><prp>/год²<n><m><nn><sg><dat>/год²<n><m><nn
1050634 ^был/быть<vbser><past><m><sg>/былой<adj><sint><short><m><sg>$
850890 ^для/для<pr>/длить<vblex><impf><tv><pprs><adv>/длить<vblex><impf><iv><pprs><adv>$
668092 ^до/до<pr>/до<n><nt><nn><sg><nom>/до<n><nt><nn><sg><gen>/до<n><nt><nn><sg><dat>/до<n><nt><
до<n><nt><nn><pl><nom>/до<n><nt><nn><pl><gen>/до<n><nt><nn><pl><dat>/до<n><nt><nn><pl><acc>/до<n>
pl><ins>$
640528 ^он/он<prn><pers><p3><m><sg><nom>/он²<n><nt><nn><sg><nom>/он²<n><nt><nn><sg><gen>/он²<n><n
он²<n><nt><nn><sg><prp>/он²<n><nt><nn><sg><ins>/он²<n><nt><nn><pl><nom>/он²<n><nt><nn><pl><gen>/о
он²<n><nt><nn><pl><ins>$
474296 ^были/быть<vbser><past><mfn><pl>/быль<n><f><nn><sg><gen>/быль<n><f><nn><sg><dat>/быль<n><f
быль<n><f><nn><pl><acc>$
472484 ^была/быть<vbser><past><f><sg>/былой<adj><sint><short><f><sg>$
460999 ^было/быть<vbser><past><nt><sg>/былой<adj><sint><cmp>/былой<adj><sint><short><nt><sg>$
434773 ^при/при<pr>/пря<n><f><nn><sg><gen>/пря<n><f><nn><pl><nom>/пря<n><f><nn><pl><acc>/переть<v
lex><impf><iv><imp><p2><sg>$
426795 ^но/но<cnjcoo>/но<n><nt><nn><sg><nom>/но<n><nt><nn><sg><gen>/но<n><nt><nn><sg><dat>/но<n><
но<n><nt><nn><sg><ins>/но<n><nt><nn><pl><nom>/но<n><nt><nn><pl><gen>/но<n><nt><nn><pl><dat>/но<n>
nn><pl><ins>$
337712 ^после/посол<n><m><aa><sg><prp>/после<adv>/после<pr>$
337153 ^же/же<part>/же<n><nt><nn><sg><nom>/же<n><nt><nn><sg><gen>/же<n><nt><nn><sg><dat>/же<n><nt
же<n><nt><nn><pl><nom>/же<n><nt><nn><pl><gen>/же<n><nt><nn><pl><dat>/же<n><nt><nn><pl><acc>/же<n>
><pl><ins>$
318197 ^под/под<pr>/под<n><m><nn><sg><nom>/под<n><m><nn><sg><acc>$
309677 ^время/время<n><nt><nn><sg><nom>/время<n><nt><nn><sg><acc>$
296390 ^области/область<n><f><nn><sg><gen>/область<n><f><nn><sg><dat>/область<n><f><nn><sg><prp>/
```

Материал проекта

```
rus     года    год     NOUN    Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
rus     года    год     NOUN    Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur
rus     года    год     NOUN    Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur
rus     году    год     NOUN    Animacy=Inan|Case=Dat|Gender=Masc|Number=Sing
rus     году    год     NOUN    Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
...
```

Спасибо за внимание!