Multi-source synthetic treebank creation for improved cross-lingual dependency parsing

Pavel Stepachev, Mariya Sheyanova
Universal Dependencies Workshop 2018 (UDW 2018)
EMNLP 2018, Brussels

Идея

- есть малоресурсные языки
- для них очень мало данных, но есть родственные!
- что если попробовать переиспользовать данные/модели больших языков для малых?

Data

Фарерский язык (CoNLL 2018 shared task)

Языки-доноры

Фарерский язык VS Скандинавские языки

Система падежей: Nom, Acc, Dat, Gen

Три грамматических рода

Синтетический грамматический строй

Система падежей: неполная

Два рода (Букмол, Шведский, Датский)

Аналитический грамматический строй

Gold Standard

1,208 предложений (10,002 токена) из Википедии

- Фарерский морфологический анализатор => lemma, POS, набор морфологических признаков
- Ручная аннотация, проверенная носителем

Трибанк доступен по ссылке.

Baseline

Делексикализованный парсинг

Идея: синтаксическая структура близкородственных языков похожа.

Как это использовать?

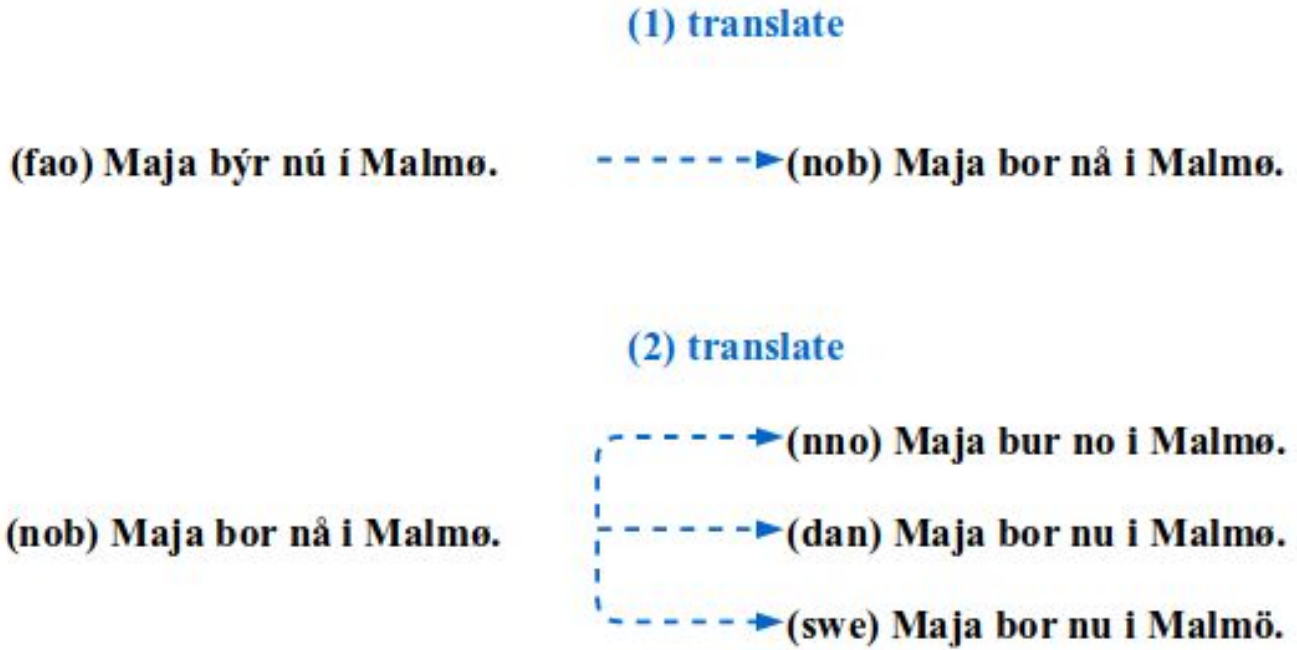
- Можно напрямую натренировать парсер на данных близких языков!
- Но в качестве фичей UDPipe использует поверхностную форму, а она может быть очень разной...
- Чтобы слова не мешали, обучаем синтаксический парсер только на POS-тэгах

Делексикализованный парсинг

Treebank	Sentences	Tokens
UD_Swedish-Talbanken	4,304	66,673
UD_Danish	4,384	80,378
UD_Norwegian-Nynorsk	14,175	245,330
UD_Norwegian-Bokmaal	15,696	243,887

Pipeline

Перевод


(1) translate
(fao) Maja býr nú í Malmø.
(nob) Maja bor nå i Malmø.
(2) translate
(nob) Maja bor nå i Malmø.
(nno) Maja bur no i Malmø.
(dan) Maja bor nu i Malmø.
(swe) Maja bor nu i Malmö.

Annotation projection

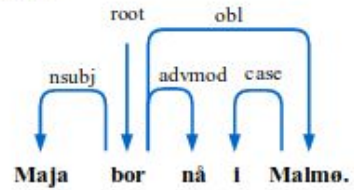
Как создавались “проекции”.

- Исходный фарерский текст и перевод выравнены пословно с помощью FastAlign (IBM Model 2)
- Переводы (dan, swe, nob, nno) проанализированы с помощью существующих для этих языков UDPipe моделей
- Парсинг переводных корпусов (все признаки, кроме токенов и лемм) перенесен обратно на фарерский

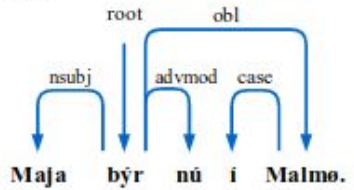
Результат: у нас есть 4 проекции (по количеству языков-доноров).

(3) project

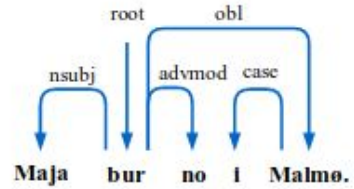
(nob)



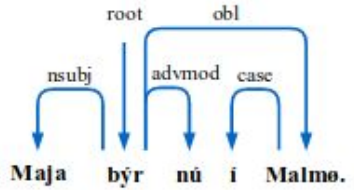
(fao)



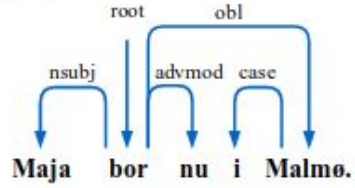
(nno)



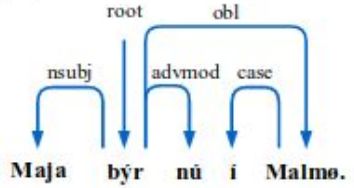
(fao)



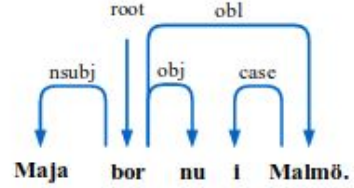
(dan)



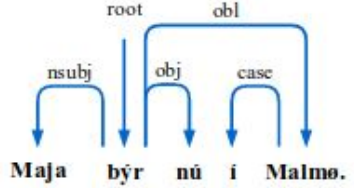
(fao)



(swe)



(fao)



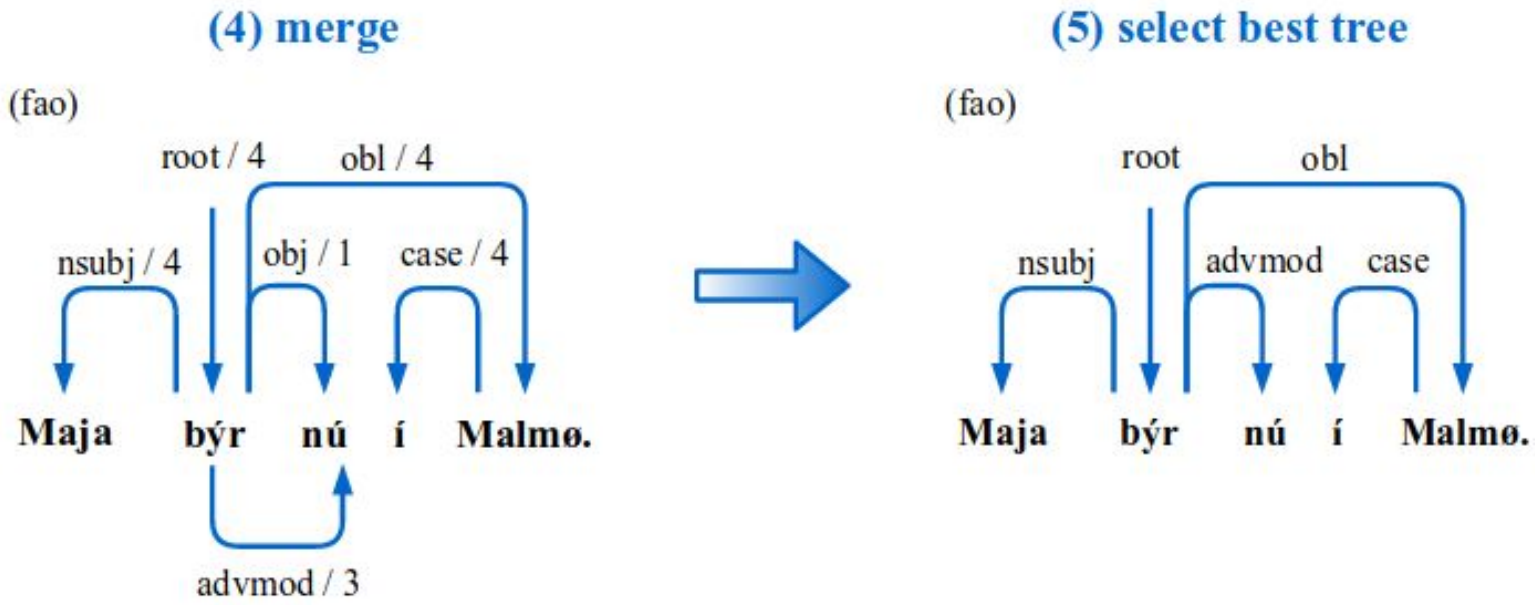
Combined model

Идея: мы хотим взять несколько (> 2) *одинаковых* UD трибанков сомнительного качества и сделать один хороший.

Как? Для каждого предложения,

- Составляем дерево с рёбрами из всех трибанков. Если одно и то же ребро встретилось 2 раза, присваиваем ему вес два, и так далее.
- Находим максимальное остовное дерево алгоритмом Chu-Liu-Edmonds.
- В результате, получаем валидные деревья, с которыми “согласна” большая часть трибанков.

Combined model



Результаты

Метрики

LAS (labeled attachment score):

количество правильных связей с учётом их типов (heads + deprels)

UAS (unlabeled attachment score):

количество правильных связей без учёта их типов (только heads)

По определению, **UAS** не может быть меньше, чем **LAS**.

Что получилось

Model	Delexicalised			Projected		
	POS	UAS	LAS	POS	UAS	LAS
Swedish	43.83	23.14	10.32	73.06	65.66	58.53
Danish	46.15	21.27	13.01	74.76	68.74	59.84
Norwegian Bokmål	44.29	24.51	15.62	**74.89**	**72.04**	**63.95**
Norwegian Nynorsk	**51.30**	**27.76**	**18.93**	72.93	70.62	62.27
Multi-source	—	—	—	**74.49**	**72.90**	**64.43**

Что хочется ещё попробовать

- другие модели выравнивания
- более умные эвристики для сопоставления one-to-many токенов
- другие системы машинного перевода: NMT
  - Yandex
  - Google
  - Microsoft
- другие языки:
  - немецкий/ идиш
  - русский/ белорусский

Что почитать? [1]

Быстро и просто:

Изучаем синтаксические парсеры для русского языка (Денис Кирьянов)

Основательно:

Speech and Language Processing, 3nd Edition (Jurafsky & Martin)

Что почитать? [2]

CoNLL

The Conference on Computational Natural Language Learning is a top-tier conference, yearly organized by SIGNLL (ACL's Special Interest Group on Natural Language Learning).

Multilingual Parsing from Raw Text to Universal Dependencies

UDW-18

The Second Workshop on Universal Dependencies invites papers on all topics relevant to universal dependencies. Priority will be given to papers that adopt a cross-lingual perspective.