Cross-Lingual Parsing

supervised cross-lingual dependency parser

(обучения парсера на дереве и применения его в целевом тексте)

Трибанки и оценка качества

Для оценки таких парсеров нам требуются как минимум следующие три компонента:

- генераторы парсеров: обучаемые, независимые от языка системы парсинга зависимостей
- деревья зависимостей для исходных языков
- оценочные наборы для целевых языков (золотой стандарт)

В 2006 и 2007 годах исследователи активно агитировали всех создавать парсеры зависимостей (Buchholz & Marsi, 2006; Nivre, Hall, Kubler, McDonald, Nilsson, Riedel, & Yuret, 2007).

Сейчас таких парсеров много, у ученых есть большой выбор, а главное -- возможность сравнивать их все между собой, что тоже сейчас является простой задачей. Таким образом, разработка кросслингвистического парсера зависимостей сводится к выбору данных для обучения, тестированию и оценке.

Внутренняя и внешняя оценка

Внутренняя оценка - оценка различных аспектов точности разбора.

Внешняя оценка - оцениваются результаты, полученные при выполнении последующих задач, где был использован данный парсер.

- labeled attachment scores (LAS) и unlabeled attachment scores (UAS): считается доля правильно соединенных вершин и зависимостей в деревьях зависимостей.

- labeled exact match scores (LEM) и unlabeled exact match scores (UEM): считается количество точных совпадений множеств, чтобы определить, как часто парсеры правильно анализируют целое предложение.

Схемы синтаксической аннотации

Схемы синтаксической аннотации обычно отличаются:

- правилами присоединения зависимых к вершинам
- метками синтаксических отношений

Дабы сократить различия в разметке деревьев зависимостей создали единую систему аннотации: Google Universal Treebanks (UDT) (McDonald и др., 2013), которая впоследствии была заменена на Universal Dependencies (UD) (Nivre et al., 2015).

Это позволило сделать первые надежные эксперименты по анализу деревьев зависимостей в разных языках и использовать метрику оценки LAS по умолчанию, как это делается при оценке одноязычных парсеров.

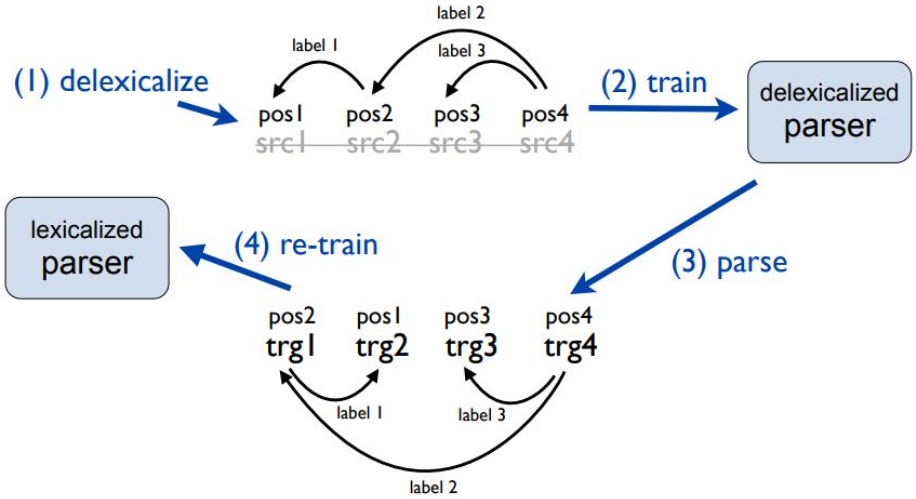
Частеречная аннотация

Парсеры зависимостей активно используют в своей работе информацию о частях речи.

Проблема совместимости частей речи, возможно, менее трудна для решения, чем структурные или маркирующие различия в деревьях зависимостей, поскольку теги частей речи более или менее прямо сопоставляются друг с другом.

Перевод модели

- Самая простая модель: применение исходной модели без всякой адаптации. Это может быть весьма успешно при близкоредственных языках.
- Деликсикализованные парсеры, которые используют UPoS теги. + можно улучшить используя несколько исходных близкородственных языков + можно перетренировать модель после делексикализации.



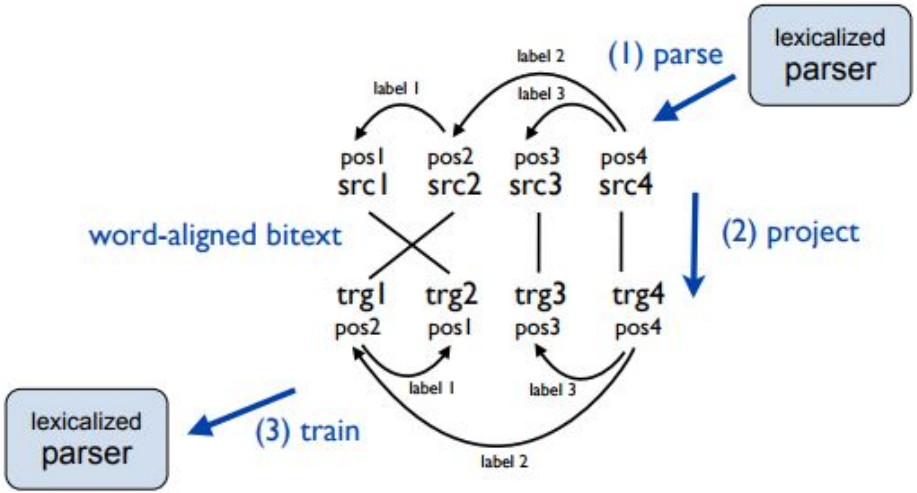
Кросс-лингвистическая репрезентация слов

Крайне важно делать выравнивание в исходном и целевом языках, а не просто использовать модель обученную на одном языке для другого языка.

Однако для этого нам необходимо наличие параллельных текстов.

Проекция аннотации

Основная идея состоит в том, чтобы использовать существующие инструменты и модели для аннотирования исходной стороны параллельного корпуса, затем использовать выравнивание, чтобы перенести отображение этой аннотации на целевую сторону корпуса. Частеречные метки обычно также переносятся вместе с отношениями между

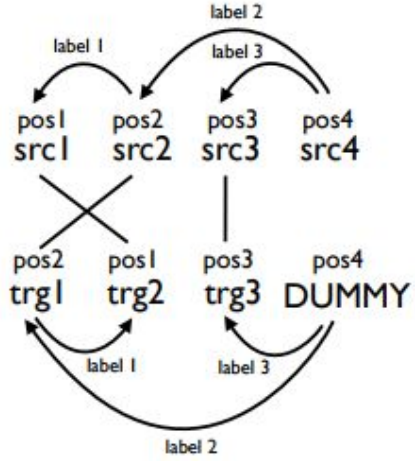


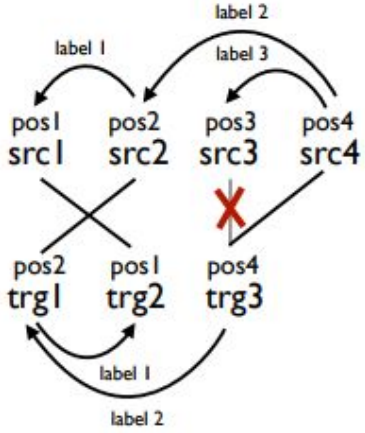
Проекция аннотации: выравнивание

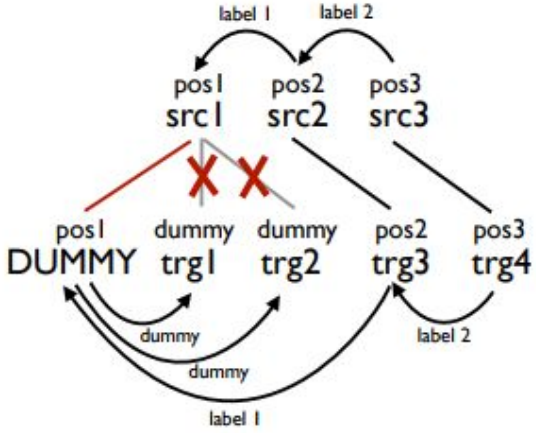
One-to-one выравнивание -- самый простой случай, когда отношения зависимостей можно просто скопировать, а неравнозначным маркерам исходного языка приписываются DUMMY узлы.

Many-to-one: сохраняется ссылка на заголовок выровненных маркеров исходного языка и удаляются все другие ссылки.

Many-to-many: Сначала мы применяем правило для выравниваний one-to-one, после чего many-to-one. Наконец, невыравненные токены целевого языка просто опускаются и затем будут вообще удалены из целевого предложения.







Перевод трибанков

Tiedemann et al. (2014) являются первыми, кто использует полномасштабный статистический машинный перевод (SMT).

1. Нужен параллельный корпус исходного и целевого языков и большой одноязычный корпус целевого языка для обучения (идеально подходящей) SMT-системы или, если можно, примените уже существующую система машинного перевода.
2. трибанк исходного языка переводится на целевой, делается выравнивание.
3. используя выравнивание, проецируется аннотация зависимостей из исходного трибанка в перевод на целевой язык.
4. тренируется парсер на данных целевого языка.

Synthetic Treebanking Experiments

Серия экспериментов, основанных на проекции аннотации и переводов трибанков.

Тренировочные данные: UDT 1

Языки: английский, французский, немецкий, испанский, шведский.

Оценка: LAS

Baseline

Делексикализированная модель в целом теряет 10 LAS в сравнении с лексикализированной моделью.

	target language →				
LAS	DE	EN	ES	FR	SV
DE	**70.84**	45.28	48.90	49.09	52.24
EN	48.60	**82.44**	56.25	58.47	59.42
ES	47.16	47.31	**71.45**	62.39	54.63
FR	46.77	47.94	62.66	**73.71**	54.89
SV	52.53	48.24	52.95	55.02	**74.55**
mate-tools (coarse)	78.38	91.46	82.30	82.30	84.52
mate-tools (full)	80.34	92.11	83.65	82.17	85.97

Улучшенная проекция аннотации

Проекция аннотации используется совместно с двуязычными word-aligned параллельными корпусами. Используется корпус Europarl. Как видно из таблицы, результаты получаются гораздо выше, чем у делексикализированных моделей.

	DE	EN	ES	FR	SV
DE	–	53.27	57.69	60.49	65.25
EN	62.28	–	62.29	65.54	66.97
ES	60.46	49.34	–	68.10	64.67
FR	61.27	53.46	66.51	–	62.75
SV	62.96	51.07	61.82	64.99	–

Улучшенная проекция аннотации

Результаты с удаленными DUMMY-узлами

	DE	EN	ES	FR	SV
DE		**53.54**$^{+0.27}$	****60.17**$^{+2.48}$	****62.35**$^{+1.86}$	****66.99**$^{+1.74}$
EN	****62.97**$^{+0.69}$		****63.80**$^{+1.51}$	****66.47**$^{+0.93}$	**67.19**$^{+0.22}$
ES	59.88$^{-0.58}$	48.85$^{-0.49}$		**68.55**$^{+0.45}$	****65.33**$^{+0.66}$
FR	**61.59**$^{+0.32}$	53.12$^{-0.34}$	**67.00**$^{+0.49}$		****64.52**$^{+1.77}$
SV	62.16$^{-0.80}$	**51.31**$^{+0.24}$	***62.58**$^{+0.76}$	**65.38**$^{+0.39}$	

Перевод трибанков на основе фраз

Плюсы:

- машинный перевод использует аналогичные синтаксические структуры, что не всегда возможно при переводе текста человеком. (это обеспечивает удачную проекцию аннотации)

Минусы:

- качество перевода.

Перевод трибанков на основе фраз

Этот метод оказывает лучший результат для некоторых пар языков в сравнении с методом проекции аннотации.

	DE	EN	ES	FR	SV
DE		$^{**}\mathbf{56.24}^{+2.70}$	$^{**}57.65^{-2.52}$	$^{**}59.06^{-3.29}$	$^{**}64.62^{-2.37}$
EN	$^{**}59.41^{-3.56}$	–	$63.76^{-0.04}$	$^{**}\mathbf{67.99}^{+1.52}$	$\mathbf{67.52}^{+0.33}$
ES	$^{**}53.94^{-5.94}$	$^{**}\mathbf{50.65}^{+1.80}$		$^{**}\mathbf{69.70}^{+1.15}$	$^{**}62.73^{-2.60}$
FR	$^{**}57.05^{-4.54}$	$^{**}\mathbf{55.69}^{+2.57}$	$^{**}\mathbf{68.66}^{+1.66}$		$^{**}62.77^{-1.75}$
SV	$^{**}58.57^{-3.59}$	$^{**}\mathbf{53.01}^{+1.70}$	$\mathbf{62.69}^{+0.11}$	$64.76^{-0.62}$	

Синтаксический перевод трибанков

Все то же что и в phrase-based +

1. Язык-источник тегируется с помощью HunPos
2. Тегированный корпус парсится с MaltParser
3. Все полученные деревья проектируются виде деревьев составляющих

Синтаксический перевод трибанков

Этот метод побеждает phrase-based практически во всех случаях.

	DE	EN	ES	FR	SV
DE		$^{**\dagger\dagger}\mathbf{58.60}^{+5.06}$	$^{**}\mathbf{61.00}^{+0.83}$	$^{**\dagger}\mathbf{63.45}^{+1.10}$	$^{**\dagger\dagger}\mathbf{67.88}^{+0.89}$
EN	$^{**}\mathbf{62.67}^{-0.30}$		$^{**\dagger}\mathbf{64.58}^{+0.78}$	$^{\dagger\dagger}\mathbf{68.45}^{+1.98}$	$^{**\dagger\dagger}\mathbf{68.16}^{+0.97}$
ES	$^{**\dagger\dagger}\mathbf{57.13}^{-2.75}$	$^{**\dagger\dagger}\mathbf{52.65}^{+3.80}$		$^{\dagger}69.37^{+0.82}$	$^{**\dagger\dagger}\mathbf{63.55}^{-1.78}$
FR	$^{**}\mathbf{61.41}^{-0.18}$	$^{**\dagger\dagger}\mathbf{56.83}^{+3.71}$	$^{\dagger}\mathbf{68.97}^{+1.97}$		$^{\dagger\dagger}62.56^{-1.96}$
SV	$^{**}\mathbf{61.73}^{-0.43}$	$^{**\dagger\dagger}52.13^{+0.82}$	$62.34^{-0.24}$	$^{\dagger}64.50^{-0.88}$	

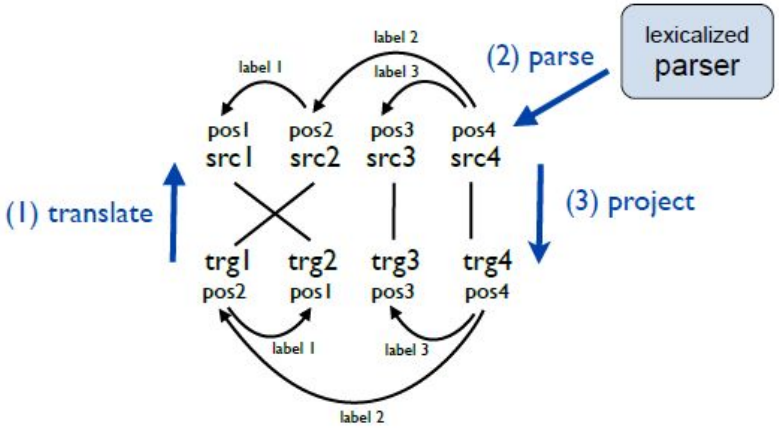
Translation and Back-Projection

Идея: Интеграция перевода в анализ.

Шаг 1. Входные данные переводятся на source language с существующими парсерами.

Шаг 2. Парсим данные на source language.

Шаг 3. Дерево разбора проецируется обратно на исходный target language.



“+”

- Предполагается,что мы доверяем парсеру (на source language), так как он натренирован на хороших трибанках.

“–”

- шум от перевода
- проблемы back-projection
  - Direct Correspondence Assumption (DCA)
- unaligned target words

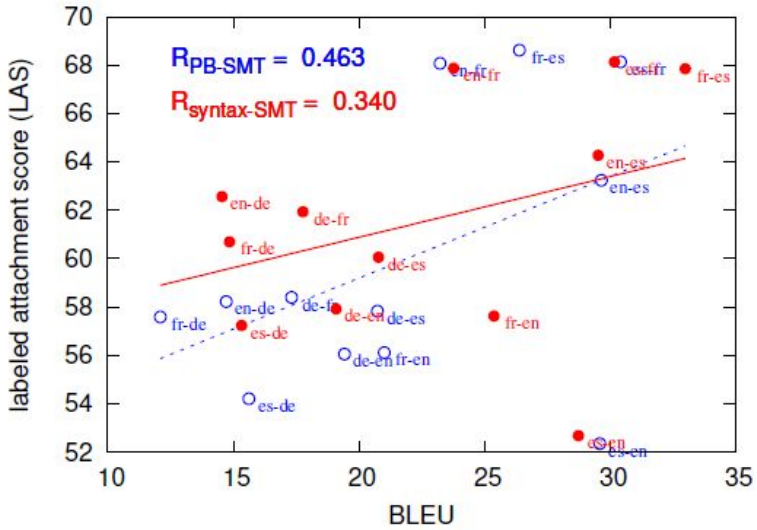
- syntax-based SMT
- ожидались результаты лучше

	DE	EN	ES	FR	SV
DE	–	$35.92^{-17.35}$	$32.90^{-24.79}$	$36.68^{-23.81}$	$45.56^{-19.69}$
EN	$44.86^{-17.42}$	–	$48.08^{-14.21}$	$48.19^{-17.35}$	$51.74^{-15.23}$
ES	$36.69^{-23.77}$	$41.91^{-7.43}$	–	$54.78^{-13.32}$	$43.23^{-21.44}$
FR	$37.44^{-23.83}$	$42.00^{-11.46}$	$55.54^{-10.97}$	–	$42.39^{-20.36}$
SV	$36.84^{-26.12}$	$35.23^{-15.84}$	$31.96^{-29.86}$	$33.74^{-31.25}$	–

Annotation Projection and Translation Quality

Существует ли корреляция между качеством перевода и производительностью cross-lingual parsers парсеров на основе переведенных трибанков?

- BLUE (WMT shared task the newstest from 2012)
- LAS



This correlation reflects the importance of the syntactic relation between languages for the success of machine translation and annotation projection.

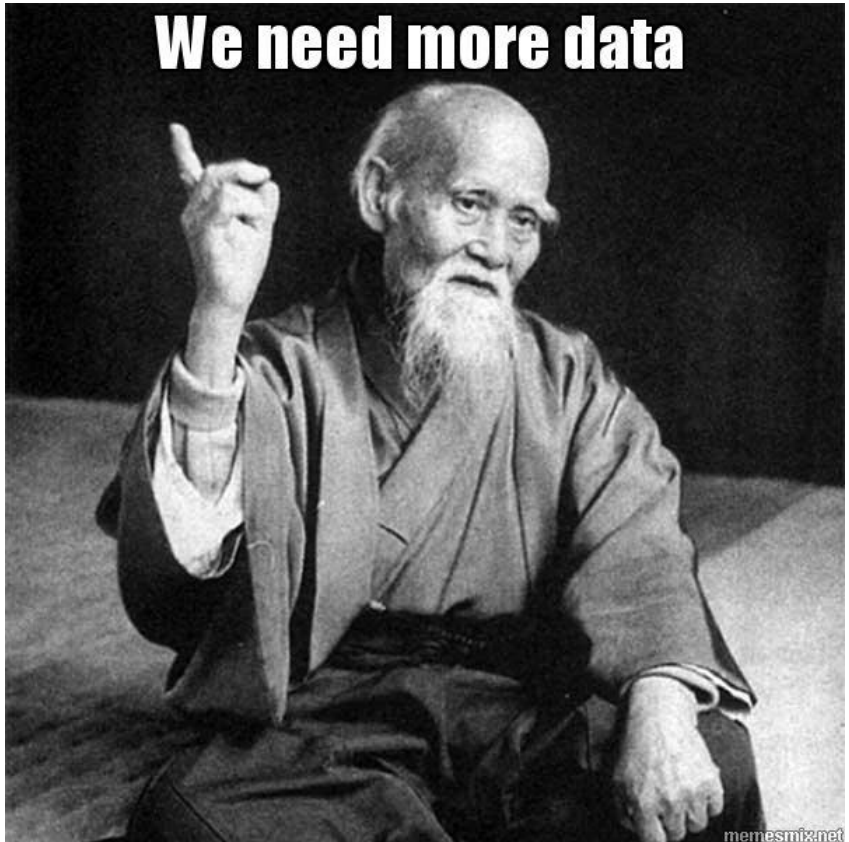
System Combination and Multi-Source Models

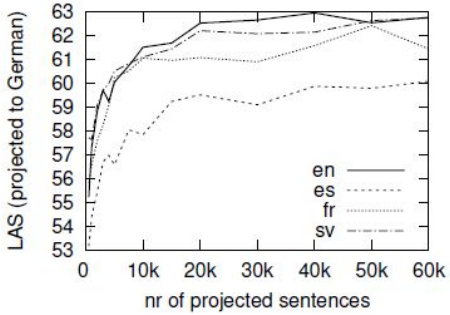
А давайте...

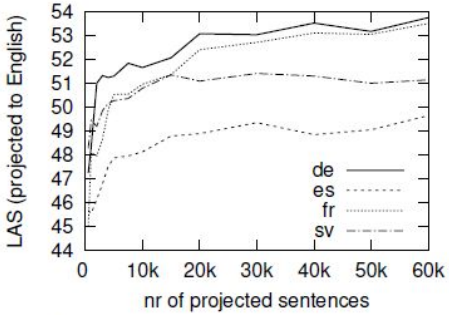
LAS	DE	EN	ES	FR	SV
best published result[4]	60.94	56.58	68.45	69.15	68.95
best individual model	63.83	58.60	68.97	69.70	68.20
annotation projection	**66.76**	55.30	*67.37*	69.48	71.95
phrase-based SMT	61.85	60.94	*68.08*	71.54	71.69
syntax-based SMT	65.89	**61.56**	***68.60***	**72.78**	**72.14**

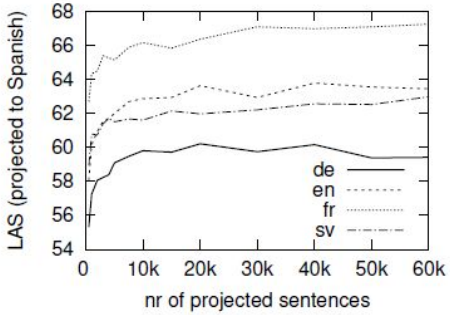
Немецкий, французский, шведский смогли приблизиться к 70%

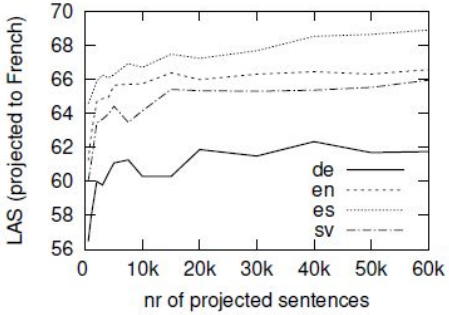
Impact of Dataset Sizes

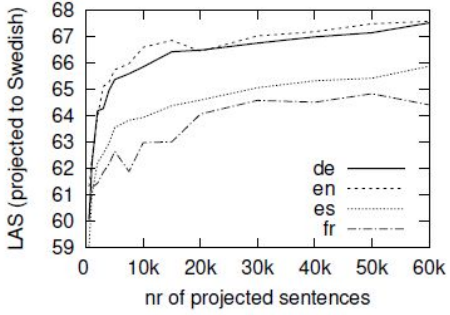



LAS (projected to German)
nr of projected sentences
en
es
fr
sv


LAS (projected to English)
nr of projected sentences
de
es
fr
sv


LAS (projected to Spanish)
nr of projected sentences
de
en
fr
sv


LAS (projected to French)
nr of projected sentences
de
en
es
sv


LAS (projected to Swedish)
nr of projected sentences
de
en
es
fr

Литература

1. Buchholz, S., & Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In Proceedings of CoNLL, pp. 149–164.
2. Nivre, J., Hall, J., K¨ubler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, pp. 915–932.
3. Jörg Tiedemann (2015) "Cross-Lingual Dependency Parsing with Universal Dependencies and Predicted PoS Labels". Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015) , pages 340–349, Uppsala, Sweden, August 24–26 2015.
4. Jörg Tiedemann and Željko Agiç (2016) "Synthetic Treebanking for Cross-Lingual Dependency Parsing". Journal of Artificial Intelligence Research 55 (2016) 209-248
5. Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering, 11 (3),311-325.

Литература

6. Michael Sejr Schlichtkrull and Anders Søgaard (2017) "Cross-Lingual Dependency Parsing with Late Decoding for Truly Low-Resource Languages". Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers

7. Mohammad Sadegh Rasooli, Michael Collins (2017) "Cross-Lingual Syntactic Transfer with Limited Resources". TACL 2017

8. Héctor Martínez Alonso, Željko Agić, Barbara Plank and Anders Søgaard (2017) "Parsing Universal Dependencies without training". Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers h