Delexicalized Parsing

Daniel Zeman, Rudolf Rosa

April 3, 2020

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Delexicalized Parsing

- What if we feed the parser with tags instead of words?
  - *Ændringer i listen i bilaget offentliggøres og meddeles på samme måde.*
  - `NNS IN NN IN NN VB CC VB IN DT NN`
  - `NNS IN NN    MD VB CC VB IN DT NN`
  - *Förändringar i förteckningen skall offentliggöras och meddelas på samma sätt.*

Delexicalized Parsing

- What if we feed the parser with tags instead of words?
  - *Ændringer i listen i bilaget offentliggøres og meddeles på samme måde.*
  - `((NNS (IN NN (IN NN))) ((VB CC VB) (IN (DT NN))))`
  - `((NNS (IN NN))    ((MD (VB CC VB)) (IN (DT NN))))`
  - *Förändringar i förteckningen skall offentliggöras och meddelas på samma sätt.*

Danish – Swedish Setup

- Daniel Zeman, Philip Resnik (2008). Cross-Language Parser Adaptation between Related Languages
  - In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pp. 35–42, Hyderabad, India

Danish – Swedish Setup

- Daniel Zeman, Philip Resnik (2008). Cross-Language Parser Adaptation between Related Languages
  - In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pp. 35–42, Hyderabad, India

- CoNLL 2006 treebanks (dependencies)
  - Danish Dependency Treebank
  - Swedish Talbanken05
- Two constituency parsers:
  - "Charniak"
  - "Brown" (Charniak N-best parser + Johnson reranker)
- Other resources
  - (JRC-Acquis parallel corpus)
  - Hajič tagger for Swedish (PAROLE tagset)

Danish – Swedish Setup

- Daniel Zeman, Philip Resnik (2008). Cross-Language Parser Adaptation between Related Languages
  - In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pp. 35–42, Hyderabad, India
- CoNLL 2006 treebanks (dependencies)
  - Danish Dependency Treebank
  - Swedish Talbanken05
- Two constituency parsers:
  - "Charniak"
  - "Brown" (Charniak N-best parser + Johnson reranker)
- Other resources
  - Hajič tagger for Swedish (PAROLE tagset)

Treebank Normalization

Danish

- DET governs ADJ
  ADJ governs NOUN

Swedish

- NOUN governs both DET and ADJ

Treebank Normalization

Danish

- DET governs ADJ
  ADJ governs NOUN
- NUM governs NOUN

Swedish

- NOUN governs both DET and ADJ
- NOUN governs NUM

Treebank Normalization

Danish

- DET governs ADJ
  ADJ governs NOUN
- NUM governs NOUN
- GEN governs NOM
  Ruslands vej
  Russia's way

Swedish

- NOUN governs both DET and ADJ
- NOUN governs NUM
- NOM governs GEN
  års inkomster
  year's income

Treebank Normalization

Danish

- DET governs ADJ
  ADJ governs NOUN
- NUM governs NOUN
- GEN governs NOM
  Ruslands vej
  Russia's way
- COORD: last member on conjunction, everything else on first member

Swedish

- NOUN governs both DET and ADJ
- NOUN governs NUM
- NOM governs GEN
  års inkomster
  year's income
- COORD: member on previous member, commas and conjs on next member

Treebank Preparation

- Transform Danish to Swedish tree style
  - A few heuristics
  - Only for evaluation! Not needed in real world.

Treebank Preparation

- Transform Danish to Swedish tree style
  - A few heuristics
  - Only for evaluation! Not needed in real world.
- Convert dependencies to constituents
  - Flattest possible structure

Treebank Preparation

- Transform Danish to Swedish tree style
  - A few heuristics
  - Only for evaluation! Not needed in real world.
- Convert dependencies to constituents
  - Flattest possible structure
- DA/SV tagset converted to Penn Treebank tags

Treebank Preparation

- Transform Danish to Swedish tree style
  - A few heuristics
  - Only for evaluation! Not needed in real world.
- Convert dependencies to constituents
  - Flattest possible structure
- DA/SV tagset converted to Penn Treebank tags
- Nonterminal labels:
  - derived from POS tags
  - then translated to the Penn set of nonterminals
- Make the parser feel it works with the Penn Treebank
  - (Although it could have been configured to use other sets of labels.)

Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)

Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)
- da-sv lexicalized: Charniak = 43.28, Brown = 41.84
  - (no morphology tweaking)

Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)
- da-sv lexicalized: Charniak = 43.28, Brown = 41.84
  - (no morphology tweaking)
- da-da delexicalized: Charniak = 79.62, Brown = 80.20 (!)
  - (hybrid sv-da Hajič-like tagset = "words", Penn POS = "tags")

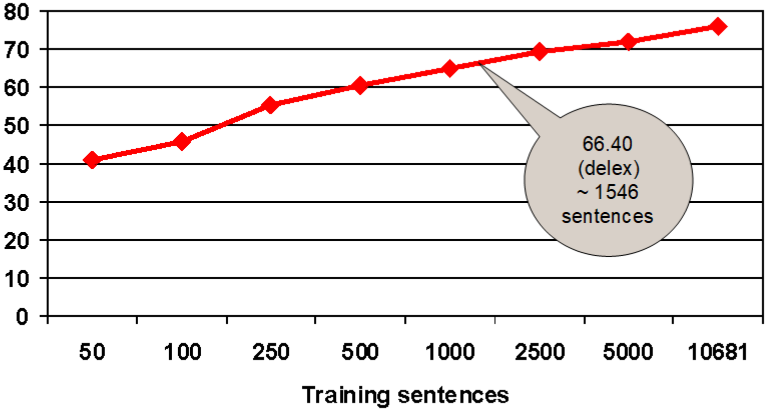
Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)
- da-sv lexicalized: Charniak = 43.28, Brown = 41.84
  - (no morphology tweaking)
- da-da delexicalized: Charniak = 79.62, Brown = 80.20 (!)
  - (hybrid sv-da Hajič-like tagset = “words”, Penn POS = “tags”)
- sv-sv delexicalized: Charniak = 76.07, Brown = 77.01

Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)
- da-sv lexicalized: Charniak = 43.28, Brown = 41.84
  - (no morphology tweaking)
- da-da delexicalized: Charniak = 79.62, Brown = 80.20 (!)
  - (hybrid sv-da Hajič-like tagset = "words", Penn POS = "tags")
- sv-sv delexicalized: Charniak = 76.07, Brown = 77.01
- da-sv delexicalized: Charniak = 65.50, Brown = 66.40

How Big Swedish Treebank Yields Similar Results?



Unlabeled F_1-score

Delexicalized Dependency Parsing

- Ryan McDonald, Slav Petrov, Keith Hall (2011). Multi-Source Transfer of Delexicalized Dependency Parsers
    - In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 62–72, Edinburgh, Scotland

Delexicalized Dependency Parsing

- Ryan McDonald, Slav Petrov, Keith Hall (2011). Multi-Source Transfer of Delexicalized Dependency Parsers
  - In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 62–72, Edinburgh, Scotland

- Transition-based parser, arc-eager algorithm, averaged perceptron, pseudo-projective technique on non-projective treebanks

Delexicalized Dependency Parsing

- Ryan McDonald, Slav Petrov, Keith Hall (2011). Multi-Source Transfer of Delexicalized Dependency Parsers
  - In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 62–72, Edinburgh, Scotland

- Transition-based parser, arc-eager algorithm, averaged perceptron, pseudo-projective technique on non-projective treebanks
- Google universal POS tags, two scenarios:
  - Gold-standard (just converted)
  - Projected across parallel corpus from English

Delexicalized Dependency Parsing

- Ryan McDonald, Slav Petrov, Keith Hall (2011). Multi-Source Transfer of Delexicalized Dependency Parsers
  - In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 62–72, Edinburgh, Scotland

- Transition-based parser, arc-eager algorithm, averaged perceptron, pseudo-projective technique on non-projective treebanks
- Google universal POS tags, two scenarios:
  - Gold-standard (just converted)
  - Projected across parallel corpus from English
- UAS (unlabeled attachment score)
- No tree structure harmonization

Delexicalized Dependency Parsing

- Ryan McDonald, Slav Petrov, Keith Hall (2011). Multi-Source Transfer of Delexicalized Dependency Parsers
  - In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 62–72, Edinburgh, Scotland

- Transition-based parser, arc-eager algorithm, averaged perceptron, pseudo-projective technique on non-projective treebanks
- Google universal POS tags, two scenarios:
  - Gold-standard (just converted)
  - Projected across parallel corpus from English
- UAS (unlabeled attachment score)
- No tree structure harmonization
  - "Danish is the worst possible source language for Swedish."

Multi-Source Transfer (McDonald et al., 2011)

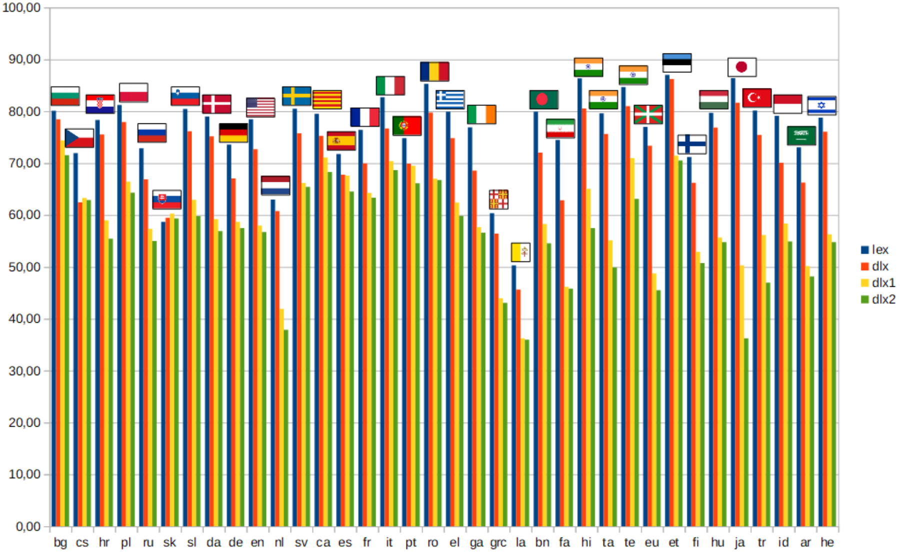
Target Test Language	Source Training Language: da	de	el	en	es	it	nl	pt	sv
da	**79.2**	45.2	44.0	45.9	45.0	<u>48.6</u>	46.1	48.1	47.8
de	34.3	**83.9**	53.2	47.2	45.8	53.4	<u>55.8</u>	55.5	46.2
el	33.3	52.5	**77.5**	<u>63.9</u>	41.6	59.3	57.3	58.6	47.5
en	34.4	37.9	<u>45.7</u>	**82.5**	28.5	38.6	43.7	42.3	43.7
es	38.1	49.4	57.3	53.3	**79.7**	<u>68.4</u>	51.2	66.7	41.4
it	44.8	56.7	66.8	57.7	64.7	**79.3**	57.6	<u>69.1</u>	50.9
nl	38.7	43.7	<u>62.1</u>	60.8	40.9	50.4	**73.6**	58.5	44.2
pt	42.5	52.0	66.6	69.2	68.5	<u>74.7</u>	67.1	**84.6**	52.1
sv	44.5	57.0	57.8	58.3	46.3	53.4	54.5	<u>66.8</u>	**84.8**

Single-Source, Harmonized (DZ, summer 2015)

- Malt Parser, stack-lazy algorithm (nonprojective)
  - Same algorithm for all, no optimization
  - Same selection of training features for all treebanks

- Trained on the first 1000 sentences only
- Tested on the whole test set
- Default score: UAS (unlabeled attachment)
- Only harmonized data used (HamleDT 3.0 = UD v1 style)
- Single source language for every target

Delexicalized Dependency Parsing with Harmonized Data



Who Helps Whom?

- Czech (62.44) ⇐ Croatian (63.27), Slovenian (62.87)
- Slovak (59.47) ⇐ Croatian (60.28), Slovenian (59.32)
- Polish (77.92) ⇐ Croatian (66.42), Slovenian (64.31)
- Russian (66.86) ⇐ Croatian (57.35), Slovak (55.01)
- Croatian (75.52) ⇐ Slovenian (58.96), Polish (55.42)
- Slovenian (76.17) ⇐ Croatian (62.92), Finnish (59.79)
- Bulgarian (78.44) ⇐ Croatian (74.39), Slovenian (71.52)

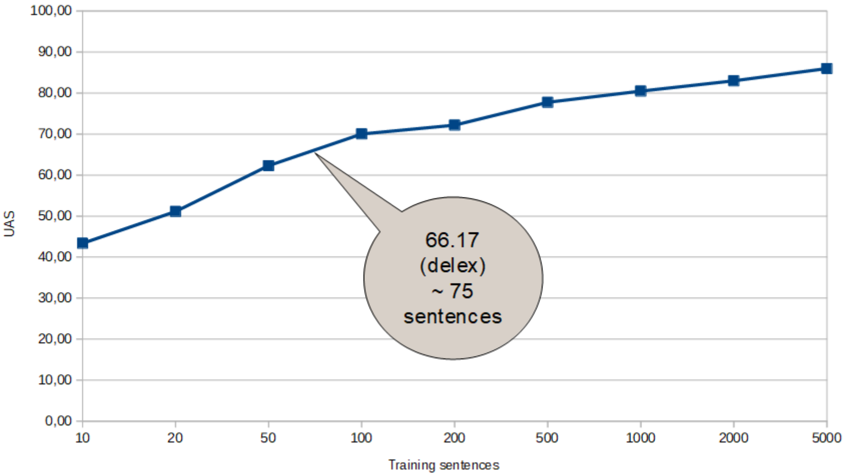
Who Helps Whom?

- Catalan (75.28) $\Leftarrow$ Italian (71.07), French (68.30)
- Italian (76.66) $\Leftarrow$ French (70.37), Catalan (68.66)
- French (69.93) $\Leftarrow$ Spanish (64.28), Italian (63.33)
- Spanish (67.76) $\Leftarrow$ French (67.61), Catalan (64.54)
- Portuguese (69.89) $\Leftarrow$ Italian (69.48), French (66.12)
- Romanian (79.74) $\Leftarrow$ Croatian (67.01), Latin (66.75)

Who Helps Whom?

- Swedish (75.73) $\Leftarrow$ Danish (66.17), English (65.41)
- Danish (75.19) $\Leftarrow$ Swedish (59.23), Croatian (56.89)
- English (72.68) $\Leftarrow$ German (57.95), French (56.70)
- German (67.04) $\Leftarrow$ Croatian (58.68), Swedish (57.48)
- Dutch (60.76) $\Leftarrow$ Hungarian (41.90), Finnish (37.89)

How Big Swedish Treebank Yields Similar Results as Delex from Danish?



Multiple Source Treebanks

- So far: select one source at a time
  - How to select the best possible source?

Multiple Source Treebanks

- So far: select one source at a time
  - How to select the best possible source?
- Alternative 1: train on all sources concatenated
  - Possibly with “weights” – take only part of a treebank, or take multiple copies of a treebank, or omit some treebanks

Multiple Source Treebanks

- So far: select one source at a time
  - How to select the best possible source?
- Alternative 1: train on all sources concatenated
  - Possibly with “weights” – take only part of a treebank, or take multiple copies of a treebank, or omit some treebanks
- Alternative 2: train on each source separately, then vote
  - Separate voting about every node’s incoming edge
  - Weights – how much do we trust each source?

Multiple Source Treebanks

- So far: select one source at a time
  - How to select the best possible source?
- Alternative 1: train on all sources concatenated
  - Possibly with "weights" – take only part of a treebank, or take multiple copies of a treebank, or omit some treebanks
- Alternative 2: train on each source separately, then vote
  - Separate voting about every node's incoming edge
  - Weights – how much do we trust each source?
  - The result should be a tree!
    - Chu-Liu-Edmonds MST algorithm, as in graph-based parsing

Multiple Source Treebanks

- So far: select one source at a time
  - How to select the best possible source?
- Alternative 1: train on all sources concatenated
  - Possibly with "weights" – take only part of a treebank, or take multiple copies of a treebank, or omit some treebanks
- Alternative 2: train on each source separately, then vote
  - Separate voting about every node's incoming edge
  - Weights – how much do we trust each source?
  - The result should be a tree!
    - Chu-Liu-Edmonds MST algorithm, as in graph-based parsing

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!
  - Problem 1: What if no relative is available? (Buryat...)

Example: CoNLL 2018 Parsing Shared Task

- Low-resource languages:
  - IE: Breton, Faroese, Naija, Upper Sorbian, Armenian, Kurmanji
  - Other: Kazakh, Buryat, Thai

Example: CoNLL 2018 Parsing Shared Task

- Low-resource languages:
  - IE: Breton, Faroese, Naija, Upper Sorbian, Armenian, Kurmanji
  - Other: Kazakh, Buryat, Thai
- High(er)-resource languages (selected groups only):
  - 1 Celtic (Irish)
  - 8 Germanic
  - 10 Slavic
  - 1 Iranian
  - 2 Turkic

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!
  - Problem 1: What if no relative is available? (Buryat…)

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!
  - Problem 1: What if no relative is available? (Buryat…)
  - Problem 2: The important characteristics may differ significantly

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!
  - Problem 1: What if no relative is available? (Buryat...)
  - Problem 2: The important characteristics may differ significantly
    - English is isolating, rigid word order

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!
  - Problem 1: What if no relative is available? (Buryat…)
  - Problem 2: The important characteristics may differ significantly
    - English is isolating, rigid word order
    - German uses morphology, freer but peculiar word order
    - Icelandic has even more morphology

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!
  - Problem 1: What if no relative is available? (Buryat…)
  - Problem 2: The important characteristics may differ significantly
    - English is isolating, rigid word order
    - German uses morphology, freer but peculiar word order
    - Icelandic has even more morphology
- WALS features (recall the first week)

Syntactic Similarity of Languages

- Observation: We cannot compare trees!
  - In real-world applications, target trees will not be available
- Language genealogy
  - Targeting a Slavic language? Use Slavic sources!
  - Problem 1: What if no relative is available? (Buryat...)
  - Problem 2: The important characteristics may differ significantly
    - English is isolating, rigid word order
    - German uses morphology, freer but peculiar word order
    - Icelandic has even more morphology
- WALS features (recall the first week)
- Language recognition tool
  - But it relies on orthography!
  - cs: *Generál přeskupil síly ve Varšavě.*
  - pl: *Generał przegrupował siły w Warszawie.*
  - ru: *Генерал перегруппировал войска в Варшаве.*
  - en: *The general regrouped forces in Warsaw.*

Measuring Treebank Similarity: POS Tag N-grams

	en	de	it	cs
DET ADJ NOUN	1.51	**1.99**	0.96	0.40
DET NOUN ADJ	0.05	0.26	**1.77**	0.10
#sent ADJ NOUN	0.13	0.09	0.02	**0.52**
NOUN PUNCT #sent	2.44	1.18	1.41	**2.73**
VERB PUNCT #sent	0.48	**1.48**	0.23	0.58

Kullback-Leibler Divergence

- $UPOS$ … universal set of 17 coarse-grained tags (from UD)
- $UPOS' = UPOS \cup \{\#sent\}$ … added sentence boundaries
- (t_{i-2}, t_{i-1}, t_i) where $t_{i-2}, t_{i-1}, t_i \in UPOS'$ … trigram of tags at positions $i-2$ … i of the corpus

Kullback-Leibler Divergence

- $UPOS$ … universal set of 17 coarse-grained tags (from UD)
- $UPOS' = UPOS \cup \{\#sent\}$ … added sentence boundaries
- (t_{i-2}, t_{i-1}, t_i) where $t_{i-2}, t_{i-1}, t_i \in UPOS'$ … trigram of tags at positions $i-2$ … i of the corpus
- $P_{Corpus}(x, y, z) = \frac{count_{Corpus}(x,y,z)}{\sum_{a,b,c \in UPOS'} count_{Corpus}(a,b,c)} = \frac{count_{Corpus}(x,y,z)}{|Corpus|}$
  - $x, y, z \in UPOS'$
  - Smoothing: need non-zero probability of every possible trigram

Kullback-Leibler Divergence

- $UPOS$ … universal set of 17 coarse-grained tags (from UD)
- $UPOS' = UPOS \cup \{\#sent\}$ … added sentence boundaries
- (t_{i-2}, t_{i-1}, t_i) where $t_{i-2}, t_{i-1}, t_i \in UPOS'$ … trigram of tags at positions $i-2$ … i of the corpus
- $P_{Corpus}(x, y, z) = \frac{count_{Corpus}(x,y,z)}{\sum_{a,b,c \in UPOS'} count_{Corpus}(a,b,c)} = \frac{count_{Corpus}(x,y,z)}{|Corpus|}$
  - $x, y, z \in UPOS'$
  - Smoothing: need non-zero probability of every possible trigram
- $D_{KL}(P_A || P_B) = \sum_{x,y,z} P_A(x, y, z) \cdot \log \frac{P_A(x,y,z)}{P_B(x,y,z)}$

Kullback-Leibler Divergence

- $UPOS$ … universal set of 17 coarse-grained tags (from UD)
- $UPOS' = UPOS \cup \{\#sent\}$ … added sentence boundaries
- (t_{i-2}, t_{i-1}, t_i) where $t_{i-2}, t_{i-1}, t_i \in UPOS'$ … trigram of tags at positions $i-2$ … i of the corpus
- $P_{Corpus}(x, y, z) = \frac{count_{Corpus}(x,y,z)}{\sum_{a,b,c \in UPOS'} count_{Corpus}(a,b,c)} = \frac{count_{Corpus}(x,y,z)}{|Corpus|}$
    - $x, y, z \in UPOS'$
    - Smoothing: need non-zero probability of every possible trigram
- $D_{KL}(P_A||P_B) = \sum_{x,y,z} P_A(x, y, z) \cdot \log \frac{P_A(x,y,z)}{P_B(x,y,z)}$
- $KL_{cpos^3}(tgt, src) = D_{KL}(P_{tgt}||P_{src})$

Kullback-Leibler Divergence

- $UPOS$ ... universal set of 17 coarse-grained tags (from UD)
- $UPOS' = UPOS \cup \{\#sent\}$ ... added sentence boundaries
- (t_{i-2}, t_{i-1}, t_i) where $t_{i-2}, t_{i-1}, t_i \in UPOS'$ ... trigram of tags at positions $i-2$ ... i of the corpus
- $P_{Corpus}(x, y, z) = \frac{count_{Corpus}(x,y,z)}{\sum_{a,b,c \in UPOS'} count_{Corpus}(a,b,c)} = \frac{count_{Corpus}(x,y,z)}{|Corpus|}$
  - $x, y, z \in UPOS'$
  - Smoothing: need non-zero probability of every possible trigram
- $D_{KL}(P_A || P_B) = \sum_{x,y,z} P_A(x, y, z) \cdot \log \frac{P_A(x,y,z)}{P_B(x,y,z)}$
- $KL_{cpos^3}(tgt, src) = D_{KL}(P_{tgt} || P_{src})$
  - Asymmetric: amount of info lost when using the source distribution to approximate the true target distribution
  - Rudolf Rosa, Zdeněk Žabokrtský (2015). KL_{cpos^3} – a Language Similarity Measure for Delexicalized Parser Transfer.
    - In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Short Papers*

How to Make the Languages More Similar?

- Lauriane Aufrant, Guillaume Wisniewski, François Yvon (2016). Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge
  - In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 119–130, Osaka, Japan.

- Transition-based parsers rely on word order
  - en: *the following question* (features: s0=ADJ, b0=NOUN)
  - fr: *la question suivante* (features: s0=NOUN, b0=ADJ)

How to Make the Languages More Similar?

- Lauriane Aufrant, Guillaume Wisniewski, François Yvon (2016). Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge
  - In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 119–130, Osaka, Japan.

- Transition-based parsers rely on word order
  - en: *the following question* (features: s0=ADJ, b0=NOUN)
  - fr: *la question suivante* (features: s0=NOUN, b0=ADJ)
- Preprocess training data
  - Reorder words
  - Remove words

How to Make the Languages More Similar?

- Lauriane Aufrant, Guillaume Wisniewski, François Yvon (2016). Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge
  - In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,* pp. 119–130, Osaka, Japan.

- Transition-based parsers rely on word order
  - en: *the following question* (features: s0=ADJ, b0=NOUN)
  - fr: *la question suivante* (features: s0=NOUN, b0=ADJ)
- Preprocess training data
  - Reorder words
  - Remove words
- How do we know?
  - Heuristics based on WALS

How to Make the Languages More Similar?

- Lauriane Aufrant, Guillaume Wisniewski, François Yvon (2016). Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge
  - In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 119–130, Osaka, Japan.

- Transition-based parsers rely on word order
  - en: *the following question* (features: s0=ADJ, b0=NOUN)
  - fr: *la question suivante* (features: s0=NOUN, b0=ADJ)
- Preprocess training data
  - Reorder words
  - Remove words
- How do we know?
  - Heuristics based on WALS
  - UPOS language model
    - Generate all permutations in window of 3 words

How to Make the Languages More Similar?

- Lauriane Aufrant, Guillaume Wisniewski, François Yvon (2016). Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge
  - In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 119–130, Osaka, Japan.

- Transition-based parsers rely on word order
  - en: *the following question* (features: s0=ADJ, b0=NOUN)
  - fr: *la question suivante* (features: s0=NOUN, b0=ADJ)
- Preprocess training data
  - Reorder words
  - Remove words
- How do we know?
  - Heuristics based on WALS
  - UPOS language model
    - Generate all permutations in window of 3 words
    - Discard non-projective subtrees; if nothing left, retain source sequence

How to Make the Languages More Similar?

- Lauriane Aufrant, Guillaume Wisniewski, François Yvon (2016). Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge
  - In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 119–130, Osaka, Japan.

- Transition-based parsers rely on word order
  - en: *the following question* (features: s0=ADJ, b0=NOUN)
  - fr: *la question suivante* (features: s0=NOUN, b0=ADJ)
- Preprocess training data
  - Reorder words
  - Remove words
- How do we know?
  - Heuristics based on WALS
  - UPOS language model
    - Generate all permutations in window of 3 words
    - Discard non-projective subtrees; if nothing left, retain source sequence
    - Score them by target-language model
    - Take the best permutation