

Transferring Word-Formation Networks Between Languages

Jonáš Vidra and Zdeněk Žabokrtský

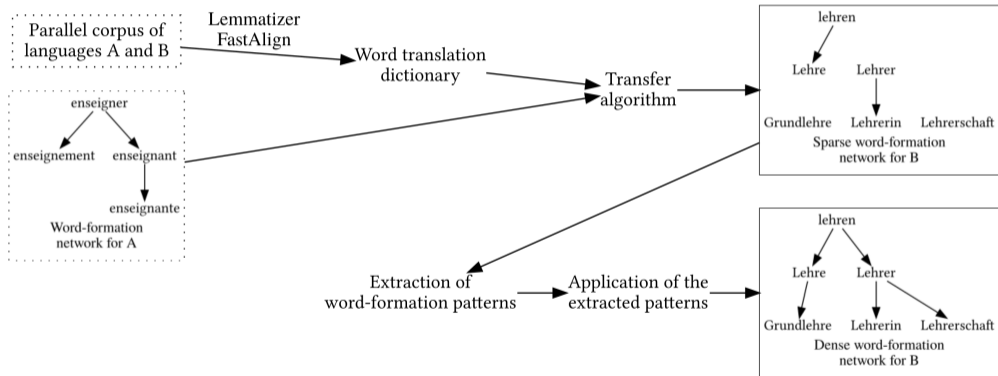
2021-09-09

Charles University

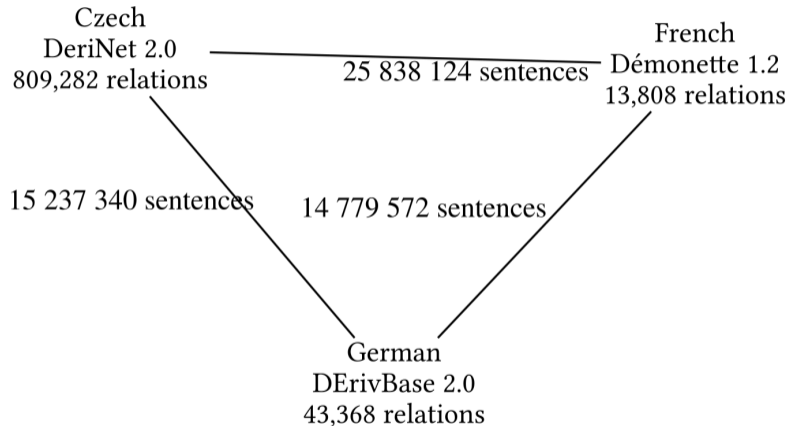
Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics

Overview



Datasets



Parallel data is from OpenSubtitles, WFNs from Universal Derivations 1.0.

Word translation lexicon induction

enseignant formateur

Lehrer Ausbilder

millier enseignant exercer métier

Tausend Lehrer Lehrerin üben Bereich Beruf

enseignant commencer expliquer matière

Lehrer beginnen Erläuterung Materie

grade instructeur ne impressionner pas

sagen doch Lehrer bedeuten gar

Hein nommer instructeur Shinsengumi

heißen nehmen Shinsengumi Lehrer Schwertkampfkunst

Word translation lexicon induction

enseignant formateur

Lehrer Ausbilder

millier enseignant exercer métier

Tausend Lehrer Lehrerin üben Bereich Beruf

enseignant commencer expliquer matière

Lehrer beginnen Erläuterung Materie

grade instructeur ne impressionner pas

sagen doch Lehrer bedeuten gar

Hein nommer instructeur Shinsengumi

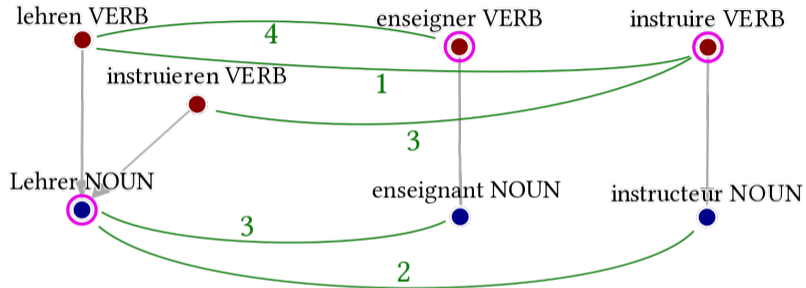
heißen nehmen Shinsengumi Lehrer Schwertkampfkunst

enseignant — Lehrer: 3

instructeur — Lehrer: 2

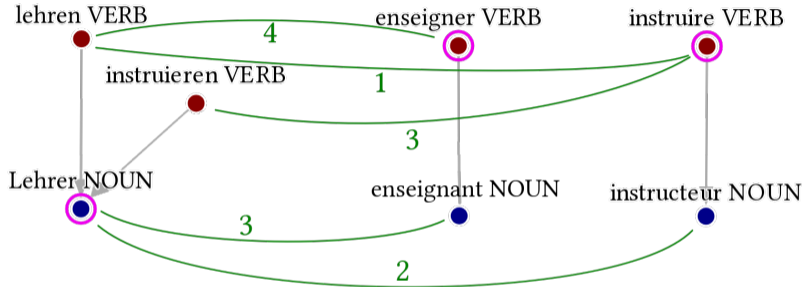
Transfer

1. Calculate transfer scores



Transfer

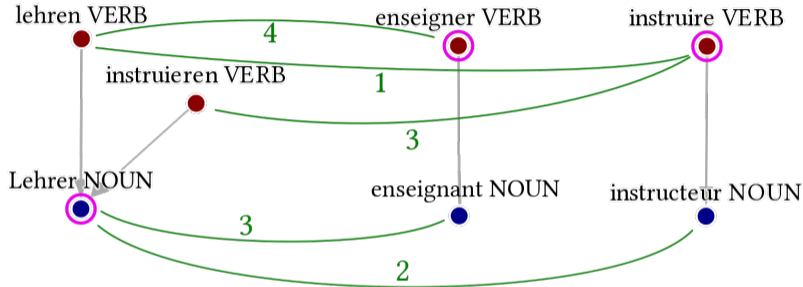
1. Calculate transfer scores



2. Calculate relative Levenshtein similarity
3. Apply threshold to the (weighted) mean of the two
4. Find maximum spanning tree

Transfer

1. Calculate transfer scores



2. Calculate relative Levenshtein similarity

3. Apply threshold to the (weighted) mean of the two

4. Find maximum spanning tree

Hyperparameters: weight of the mean, threshold.

Recall issues with the transfer

školit (“to teach”) → *školení* (“a course”)

kupit (“to heap”) → *kupení* (“a heaping”)

Recall issues with the transfer

školit (“to teach”) → *školení* (“a course”) ✓

kupit (“to heap”) → *kupení* (“a heaping”) ✗

Recall issues with the transfer

školit (“to teach”) → *školení* (“a course”) ✓

kupit (“to heap”) → *kupení* (“a heaping”) ✗

* *Bruder* (“brother”) → *Cousin* (“cousin”), due to Czech *bratr* → *bratranec*

Kampf (“a fight”) → *kämpfen* (“to fight”)

Kampf (“a fight”) → *kämpfen* (“to fight”)

Kampf (“a fight”) → kämpfen (“to fight”): affixal pattern *ka-* → *kä-* + *-en*

Expansion via supervised ML

Kampf (“a fight”) → *kampfen* (“to fight”): affixal pattern *ka-* → *kä-* + *-en*

Features:

- Frequency of the affixal pattern of the proposed relation in the training data
- POS categories of both lexemes
- Edit distance of lemmas
- Difference of lemma lengths
- Capitalization of first characters

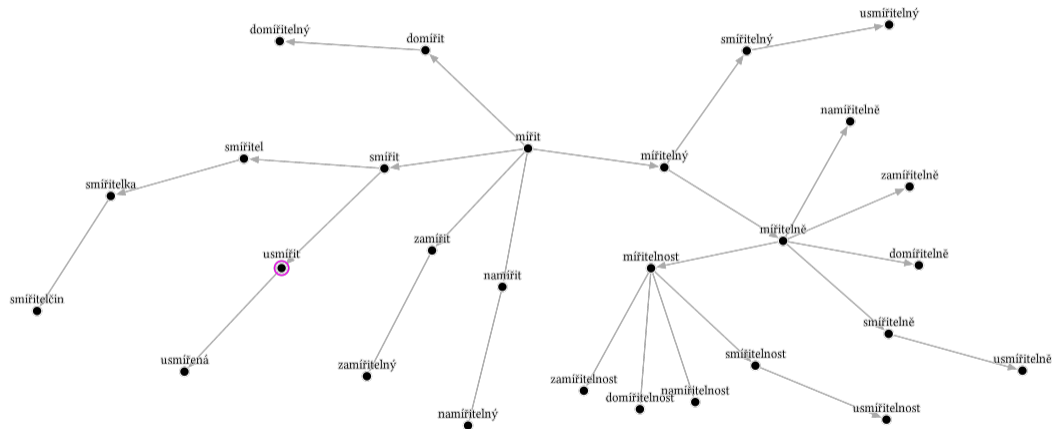
Evaluation of transfer

Lang pair	Gold scores [%]				Internal scores [%]		
	Prec.	Recall	F1	Acc.	Recall	F1	Acc.
de → cs	39.66	0.29	0.58	1.19	33.11	36.09	53.71
fr → cs	42.46	0.37	0.73	1.33	36.11	39.03	53.79
cs → de	27.06	2.45	4.50	17.07	35.36	30.66	65.88
fr → de	14.33	0.20	0.39	4.19	14.14	14.24	64.74
cs → fr	23.54	2.10	3.86	7.65	30.50	26.57	42.72
de → fr	3.47	0.04	0.07	1.84	11.36	5.32	59.45

Evaluation of supervised ML

Lang pair	Prec.	Recall	F1	Acc.
de → cs	45.70	73.81	56.45	48.90
fr → cs	39.60	70.00	50.58	43.99
cs → de	35.02	67.73	46.17	80.15
fr → de	44.25	39.35	41.66	84.62
cs → fr	60.33	88.64	71.79	66.30
de → fr	35.57	13.79	19.88	36.69

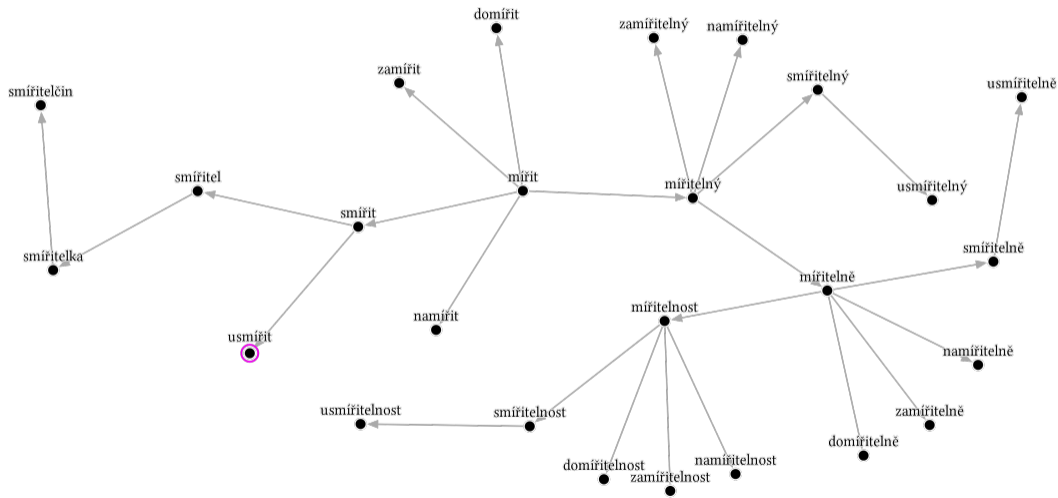
Output samples: de-cs



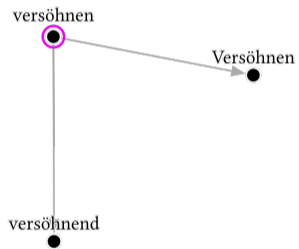
réconcilier



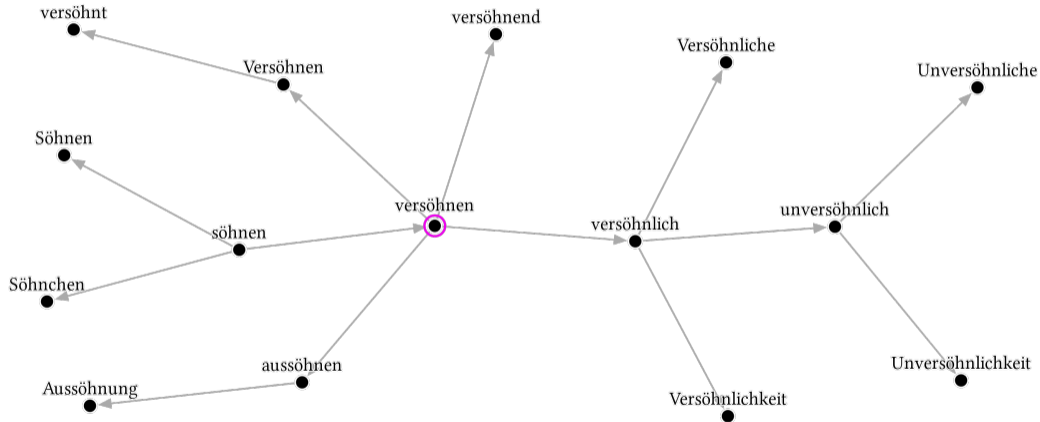
Output samples: fr-cs



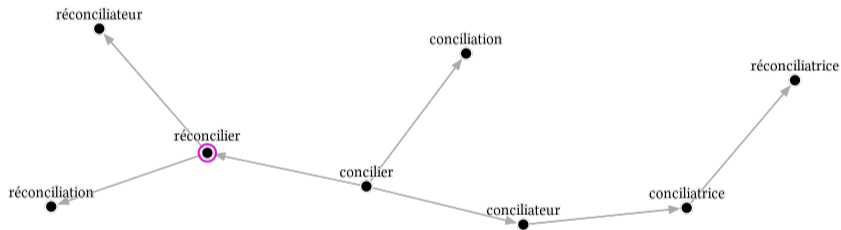
Output samples: fr-de



Output samples: cs-de



Output samples: cs-fr



Summary

- Transferring word-formation networks cross-lingually works
- The created networks are sparse
- Combining transfer with a bootstrapping step makes recall better
- Precision is not great, not terrible