Cross-lingual and supervised models for morphosyntactic annotation: a comparison on Romanian

Lauriane Aufrant[1,2], *Guillaume Wisniewski*[1], *François Yvon*[1]

(1) LIMSI-CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, (2) DGA

{lauriane.aufrant, guillaume.wisniewski, francois.yvon}@limsi.fr

Summary

We experiment with state-of-the-art cross-lingual methods in a realistic low-resource scenario, with a study on source selection and multi-source use.

↪ Results confirm intuitions on sources, but also show that their extracted amount of linguistic knowledge remains low compared to supervision.

Cross-lingual transfer

Large resources only for a few languages, poor performance for all other ones

↪ Key idea: use knowledge from well-resourced languages to build tools for low-resourced languages

Annotation projection

- Annotate the source side of parallel data
- Align words and assume that linked words share their label
- Train on that newly annotated target data

Direct delexicalized transfer

- Replace source tokens with their part-of-speech tags
- Train a source delexicalized model
- Use directly on target

Romanian as a case study

About Romanian

- Romance SVO language, with rich morphology
- BALRIC-LING, Europarl, recent efforts... but still under-resourced

Experiments

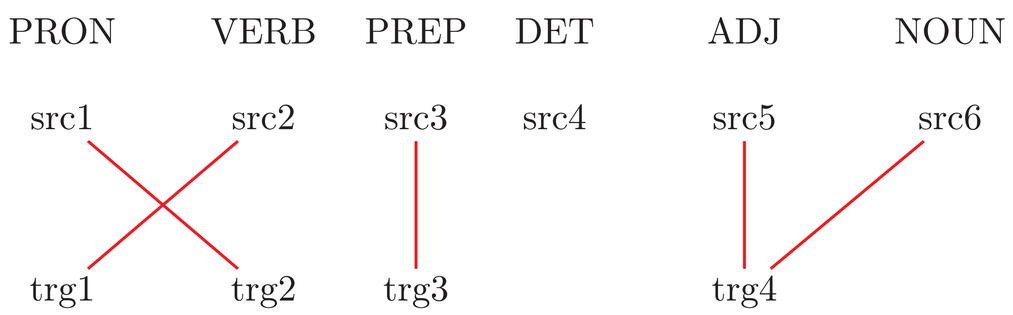
Train: Universal Dependency Treebank + Europarl

↪ Test: Romanian Syntactic Annotated Corpus [Perez, 2012]

PoS transfer

Work by [Wisniewski et al., 2014]:

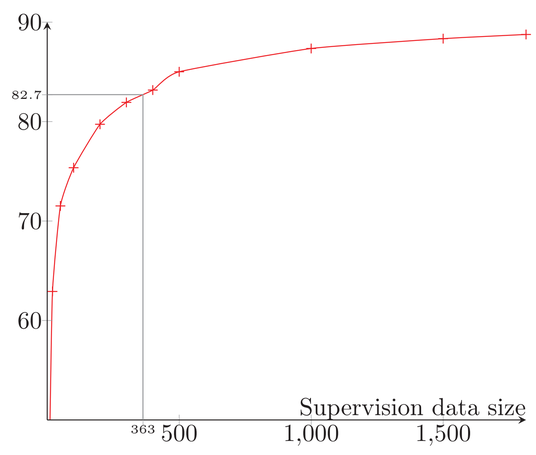
projection with ambiguous learning + crawled lexicon constraints



~~VERB~~ AUX PRON PREP ADJ/NOUN

Results (PoS accuracy)

Source	Accuracy
en	82.0
fr	82.7
it	81.8
es	82.7
fr+it+es	82.5
avg(fr,it,es)	82.4
supervised	88.8



Experiments confirm that:

- A Romance source is preferable to English.
- Multi-source is preferable to choosing a random source.

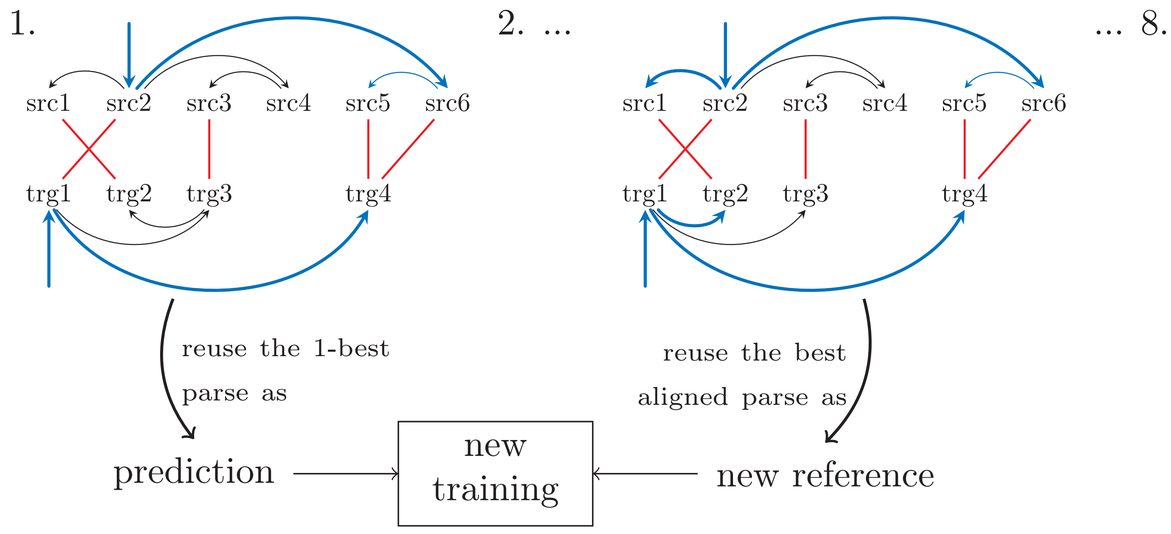
⇒ Drawback: **363** annotated sentences already reach those scores.

Dependency transfer

Inspired by [McDonald et al., 2011]:

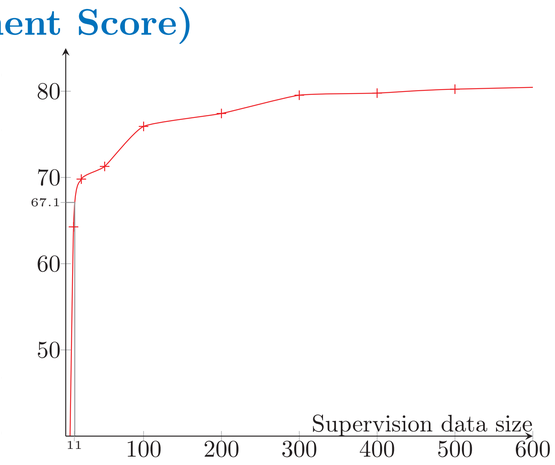
(a) seeding with delexicalized model transfer + relexicalization

(b) n-best reranking, based on parallel parse agreement



Results (Unlabeled Attachment Score)

Source	Delex	Relex	Full
en	55.6	57.4	65.7
fr	60.8	61.8	67.0
it	61.5	62.1	66.9
es	61.2	62.1	67.1
fr+it+es	61.7	61.6	67.1
avg(fr,it,es)	61.2	62.0	67.0
supervised		82.7	



⇒ Drawback: **11** annotated sentences already reach those scores.

Analysis and conclusions

Romance transfer to Romanian: regular divergences

- Annotation conventions — *să* → CONJ, PREP or PART?
- Low 1:1 token correspondence — definite article ↔ noun inflection
- Easy but target-specific rules — completives vs subordinates
  Quiero comer [es] / *Vreau să mănânc* [ro]
- Loanwords from unrelated languages — 20% Slavic adverbs
- Over-reliance on alignments — *Bob likes Mary* / *Mary pleases Bob*

Guidelines for new cross-lingual systems

- Leverage all the available target data: foster non-regressive transfer
- Finer multi-source combination with phrase-level source weighting
- Include other knowledge sources, e.g. lexical similarity
  casa: NOUN [es] ⇒ *casă*: NOUN [ro]

Our contributions to open resources

Romanian Wiktionary tag lexicon

↪ a lexicon with PoS and morphological annotations, crawled from `ro.wiktionary.org` and `en.wiktionary.org`

PanParser

↪ a modular implementation of a transition-based dependency parser, easy to customize for research purpose