基于翻译的无监督跨语言迁移学习

张梅山

天津大学新媒体与传播学院

CCF-NLP 走进华南理工大学(2020.9.6)

Cross-Lingual Transfer

Source Language (English) resource-rich

Target Language (Portuguese) resource-pool

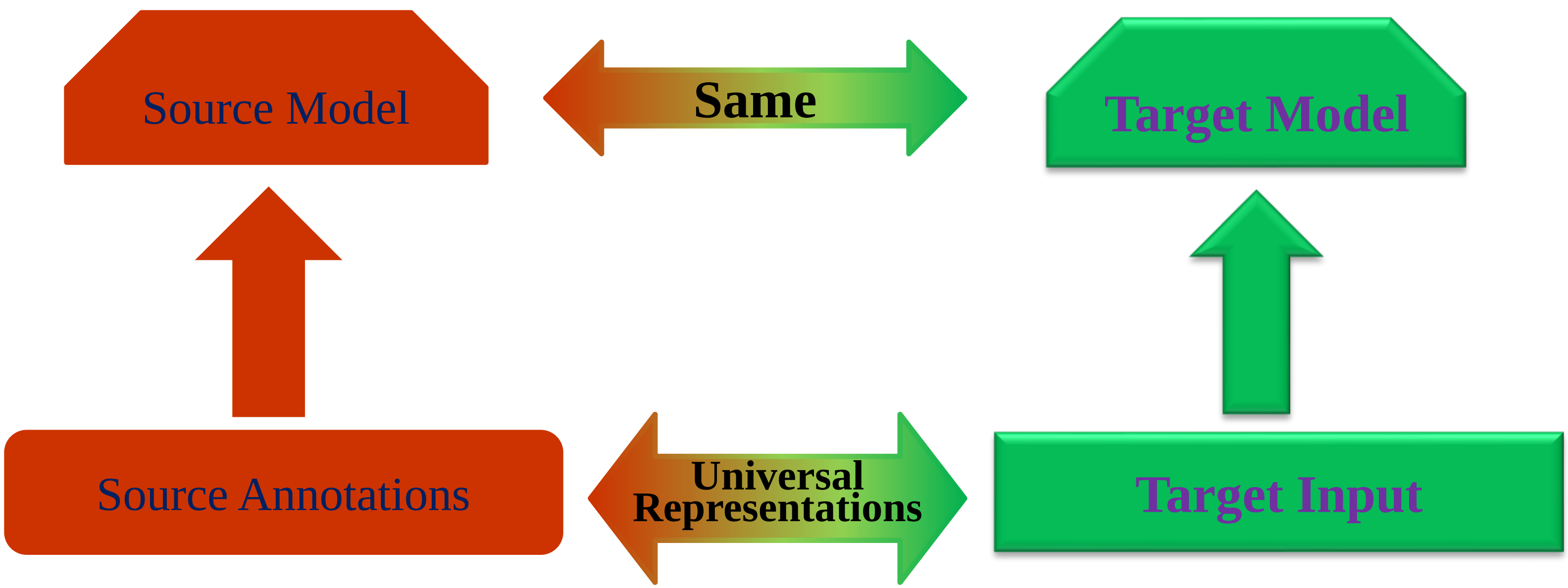
Well-trained models
with large annotations

How to supervise?

Cross-Lingual Transfer

Model Transfer

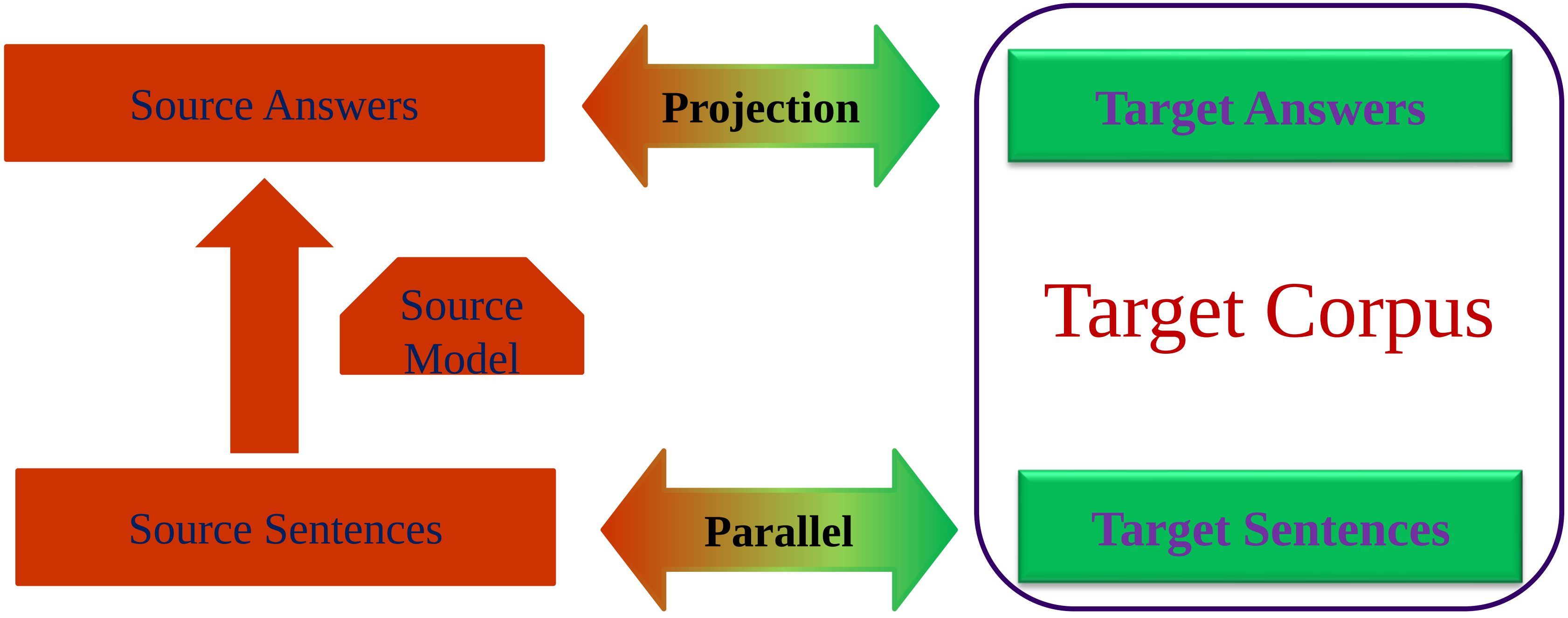
- Building cross-lingual models on language-independent features,
- such as cross-lingual word representations, universal POS tags which can be transferred into target languages directly.



Cross-Lingual Transfer

Annotation Adaptation

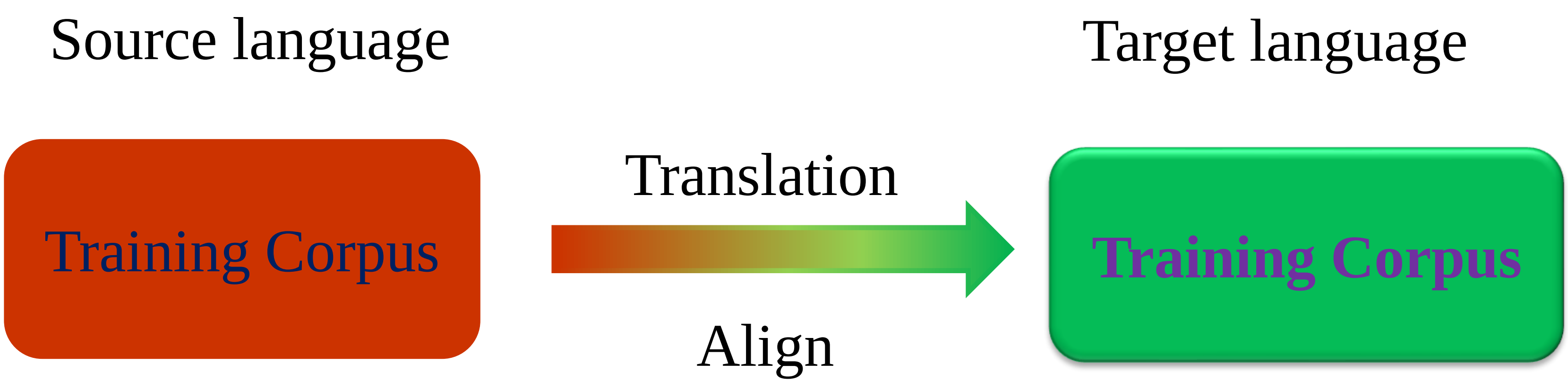
- Based on a large-scale parallel corpus between the source and target languages
- The source-side sentences are annotated with SRL tags automatically by a source SRL labeler
- The source annotations are projected onto the target-side sentences in accordance of word alignments.



Cross-Lingual Transfer

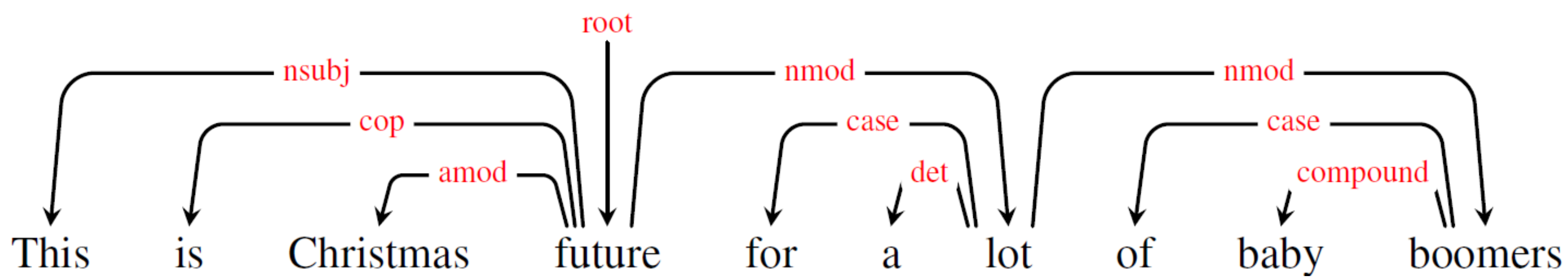
Translation-based (Our method)

- Translate the annotated sentences into target languages.
- Align the annotated information.
- Train a target model for a specific task.

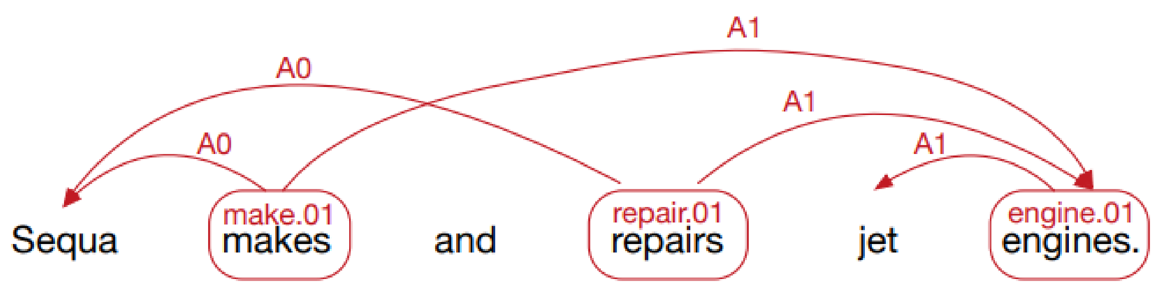


Two Tasks

- Dependency Parsing (EMNLP 2019)



- Sematic Role Labeling (ACL 2020)



Dependency Parsing

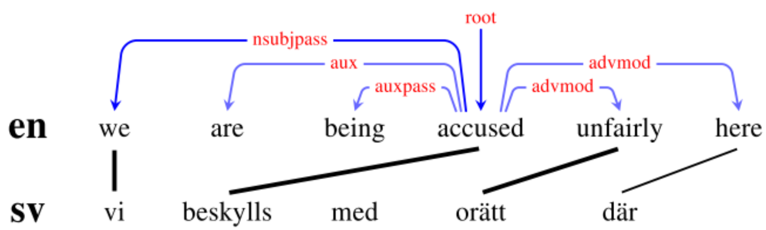
TreeBank Translation

- Tiedemann et al. (2014) EMNLP
- Tiedemann (2015)
- Tiedemann and Agic´ (2016) JAIR

Dependency Parsing

Our idea

- Derive Code-Mixed Treebank by partial translation to minimum noise
- Transfer knowledge between languages by Cross-lingual word representations



(a) full-scale translation.

(b) this method, partial translation.

from English (en) to Swedish (sv)

Meishan Zhang, Yue Zhang and Guohong Fu. Cross-Lingual Dependency Parsing Using Code-Mixed TreeBank. EMNLP 2019

Dependency Parsing

Partial Translation

- Hyper-parameter λ to control the ratio of translation

- $\lambda = 0$, no source word is substituted or deleted

  it is a source language dependence tree

- $\lambda = 1$, all words are target language

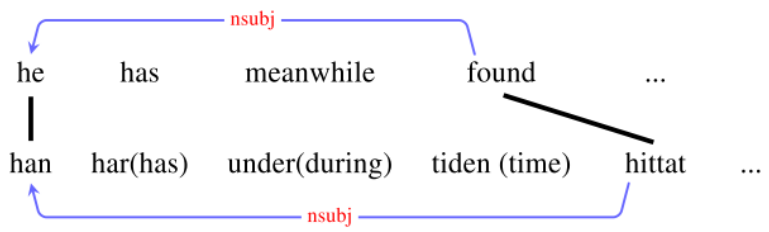
  it is a source language dependence tree

Dependency Parsing

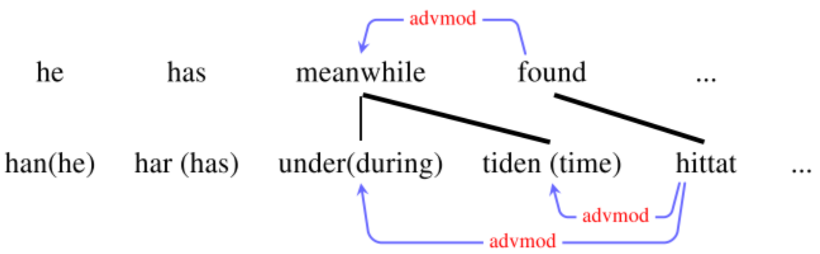
Word Substitution

- Substitutes the source words with the target translations

  source → target



(a) one-to-one.



(b) many-to-one.

Dependency Parsing

Word Substitution

- For each target word f_j , obtain the most confidently aligned source word e_i by their alignment probability

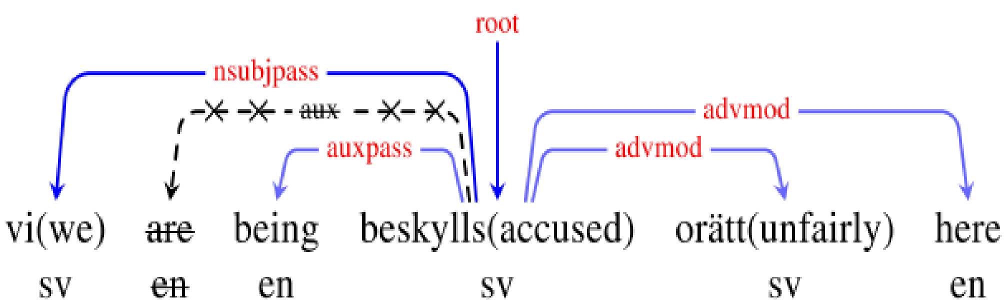
$$p_j = p(e_i|f_j)$$

- Sort the target words by p_j, choosing the top $\lceil m\lambda \rceil$ words with highest alignment probabilities for substitution (m is the length of target sentence)

Dependency Parsing

Word Deletion

- Removes several unaligned source words



Dependency Parsing

Word Deletion

- For each source word e_i who has no aligned target word, accumulate the probabilities

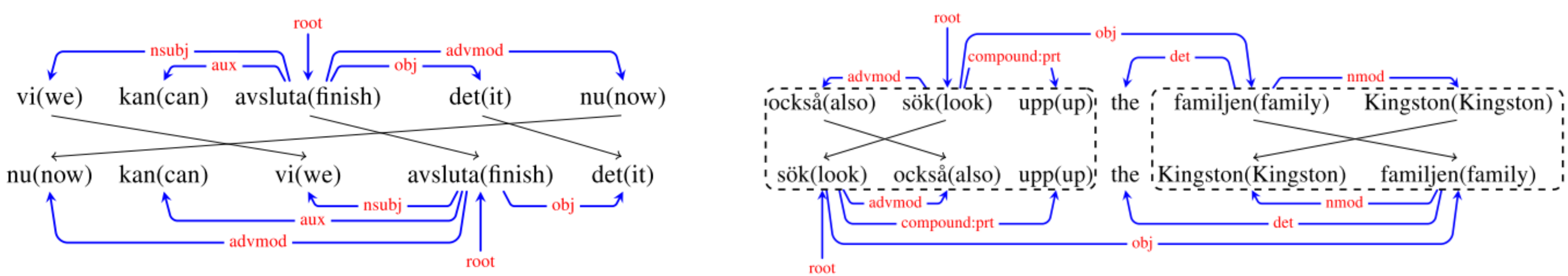
$$r_i = \sum_j p(e_i|f_j)$$

- Sort the source words by r_i, choosing the top $\lceil k\lambda \rceil$ words with lowest score for deletion (k is the num of these source words)

Dependency Parsing

Sentence Reordering

- Reorder the partially translated sentence



(a) full sentence

(b) two spans

Dependency Parsing

Sentence Reordering

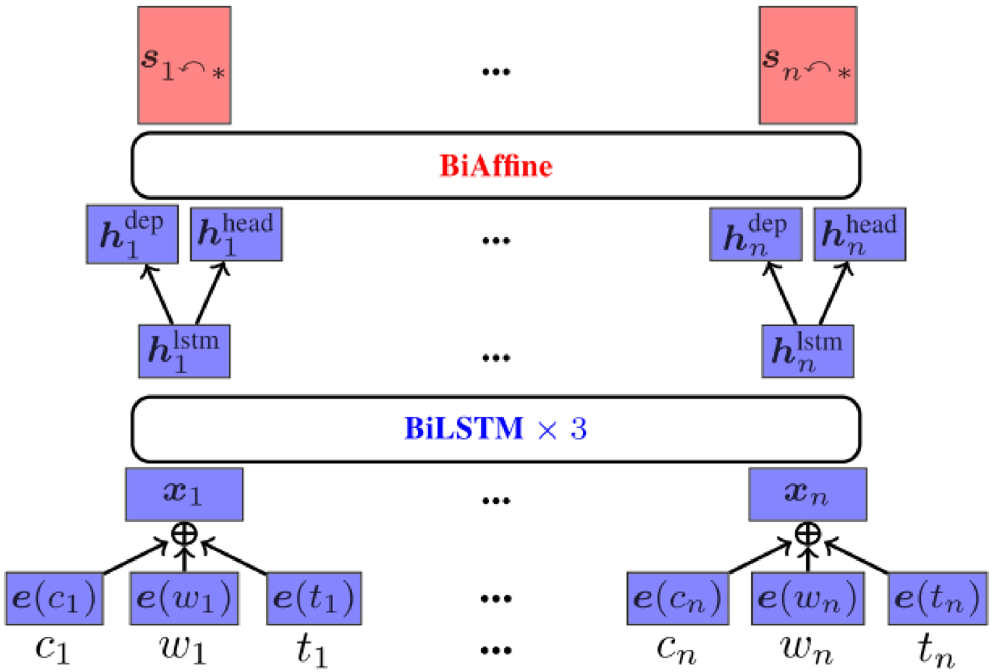
- Continuous target spans are reordered to make the final sentence contain grammatical phrases

- Continuous spans of target words interrupted by source word are reordered according to their order in the machine translation

Dependency Parsing

Parser Model

- BiAffine Parser From Dozat and Manning (2016)



Dependency Parsing

Parser Model

- **Input:**

$$x_n = e(w_n) \oplus e(c_n) \oplus e(t_n)$$

Word, Cluster and POS tag embedding

- **BiLSTM:**

$$h_n^{lstm} = BiLSTM\ (x_n)$$

- **MLP:**

$$h_n^{dep} = MLP^{dep}(h_n^{lstm})$$

$$h_n^{head} = MLP^{head}(h_n^{lstm})$$

- **Output:**

$$s_{i \frown j} = BiAffine(h_i^{dep}, h_j^{head})$$

Dependency Parsing

Datasets

- **Universal Dependency Treebanks (v2.0)**
  - Source language: English (EN)
  - Six target language: Spanish (ES), German (DE), French (FR), Italian (IT), Portuguese (PT) and Swedish (SV)
- **Google Translate as machine translation system**

Dependency Parsing

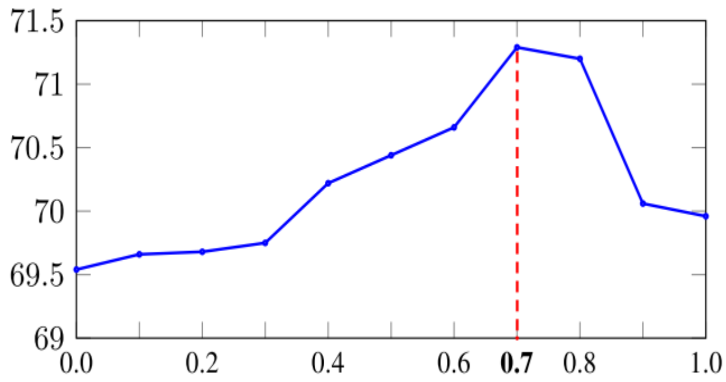
Experiment Settings

- **Delex (McDonald et al., 2013):**
  Model without cross-lingual word representations
- **PartProj (Lacroix et al., 2016):**
  Model trained on the corpus by projecting only the source dependencies
- **Src (Guo et al., 2015):**
  Model trained on the Source English treebank
- **Tgt (Tiedemann and Agi´c, 2016):**
  Model trained on the fully translated Target treebank
- **Mix (Ours):**
  Model trained on the Code-Mixed treebank

Dependency Parsing

Influence of The Translation Ratio λ

- Performance improves after translating, demonstrating the effectiveness of syntactic transferring, and reaches the peak when $\lambda = 0.7$
- There is a significant drop when λ grows from 0.8 to 0.9 because the newly added dependency are mostly noisy



Dependency Parsing

Mixing with Source TreeBank

Model	UAS	LAS
Src	79.52	69.54
Tgt	79.34	69.96
Mix	**80.33**	**71.29**
Src + Tgt	80.12	71.16
Src + Mix	**80.91**	**71.73**

- Source treebank is complementary with the translated treebanks, Src + mix gives the best performance
- But its improvement over mix is smaller than that of src+tgt over tgt, because mix contains relatively more source than the fuly translated target treebank

Dependency Parsing

Ablation Studies

Model	UAS	LAS
`Mix`	**80.33**	**71.29**
−Sentence Reordering	79.79	70.47
−Word Deletion	79.82	70.64
−Both	79.46	69.59

- Word Deletion and Sentence Reordering are important
- Without both, the performance is only comparable with the baseline

Dependency Parsing

Lang.	Delex		PartProj		Src		Tgt		Src + Tgt		Mix		**Src + Mix**	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
DE	64.10	53.77	69.90	61.28	66.87	57.46	70.84	62.30	72.41	63.74	71.41	63.46	**72.78**	**64.38**
ES	71.53	63.33	75.81	66.83	75.63	65.85	76.49	67.39	77.00	67.95	81.18	71.80	**81.44**	**71.66**
FR	75.13	67.26	75.54	67.63	78.13	70.63	76.91	69.39	78.75	71.17	83.20	76.32	**83.77**	**76.48**
IT	77.71	69.27	77.71	69.27	81.11	72.83	79.30	71.65	81.56	74.09	85.30	77.43	**86.13**	**78.38**
PT	74.03	67.70	79.44	71.30	77.37	69.36	78.32	70.67	79.73	71.84	83.54	75.34	**84.05**	**75.89**
AVG	72.50	64.27	75.68	67.26	75.82	67.23	76.37	68.28	77.89	69.76	80.93	72.87	**81.63**	**73.36**

Final Results

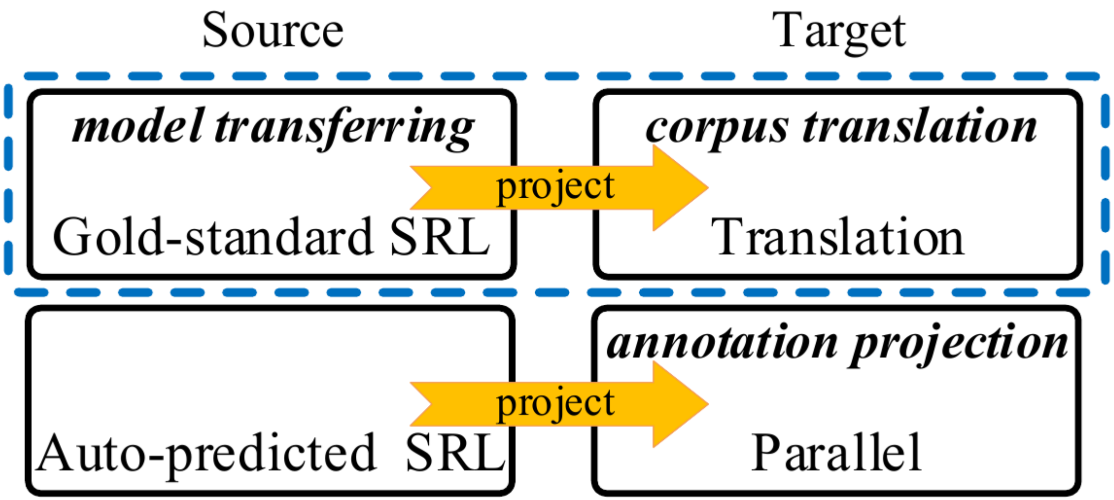
Dependency Parsing

Model	DE	ES	FR	IT	PT
TreeBank Transferring					
This	72.78	**81.44**	**83.77**	**86.13**	**84.05**
Guo15	60.35	71.90	72.93	–	–
Guo16	65.01	79.00	77.69	78.49	81.86
TA16	**75.27**	76.85	79.21	–	–
Annotation Projection					
MX14	74.30	75.53	70.14	77.74	76.65
RC15	**79.68**	**80.86**	**82.72**	**83.67**	**82.07**
LA16	75.99	78.94	80.80	79.39	–
TreeBank Transferring + Annotation Projection					
RC17	**82.1**	**82.6**	**83.9**	**84.4**	**84.6**

Comparison with Previous Work

Semantic Role Labeling

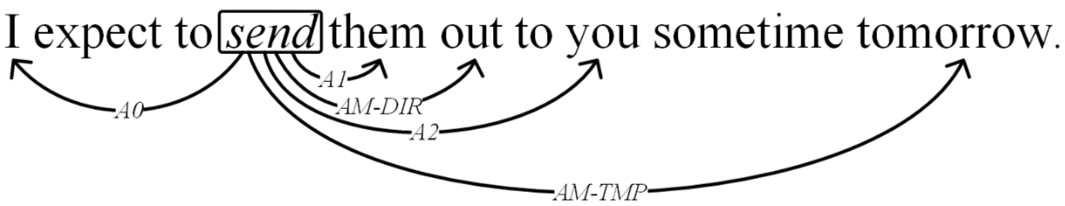
- Our method:
  - Model transfer
  - Corpus translation



Hao Fei, Meishan Zhang and Donghong Ji. Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus. ACL 2020

Semantic Role Labeling

shallow semantic parsing, recognizing the predicate-argument structure, such as:
who did what to whom, where and when, etc.



Dependency-based: **head of arguments**

- Predicate: *send*
- Argument:

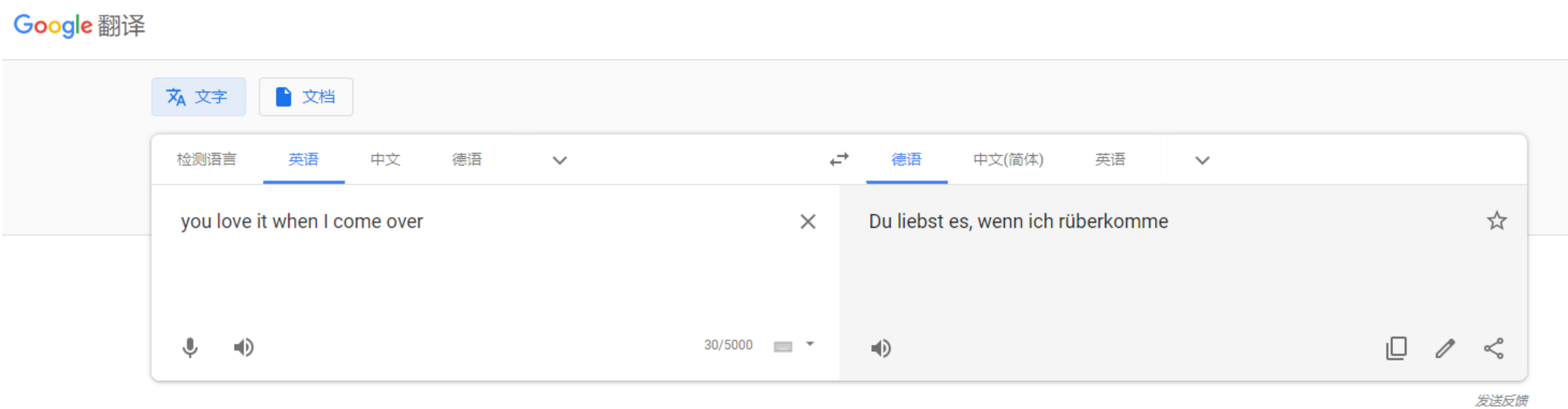
  Core roles: *I*(A0), *them*(A1), *you*(A2)

  Modifying roles: *out*(AM-DIR), *tomorrow*(AM-TMP)

Semantic Role Labeling

Step1: Translating

Source sentence $\xrightarrow{\textit{Translation System}}$ Target sentence



Semantic Role Labeling

Step2: Projecting

alignment probabilities: $a(f_j|e_i)$

POS tag distributions: $p(t_*|f_j)$

Confidence score of the projection:

$$\text{score}(e_i \rightarrow f_j, r_{e_i}) = a(f_j|e_i)p(t_{e_i}|f_j)$$

Semantic Role Labeling

◆Projection rules:

- Fine-grained:

Ideal:

(a) one-one target-source alignment.

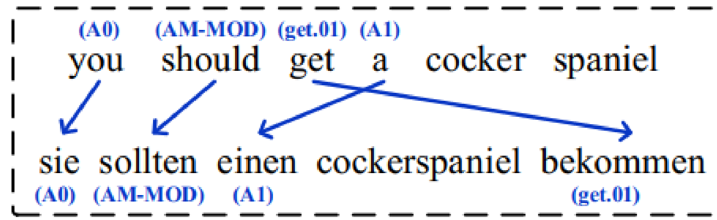
Problematic :

(b) predicate-argument collision, keep only the predicate,

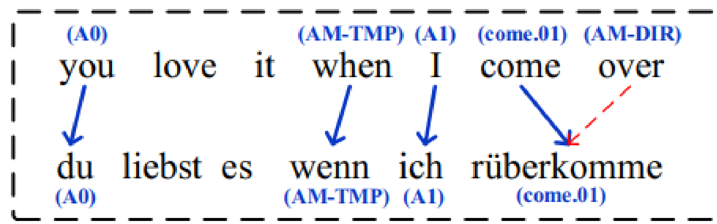
(c) argument-argument collision, keep the one with higher confidence.

- Overall:

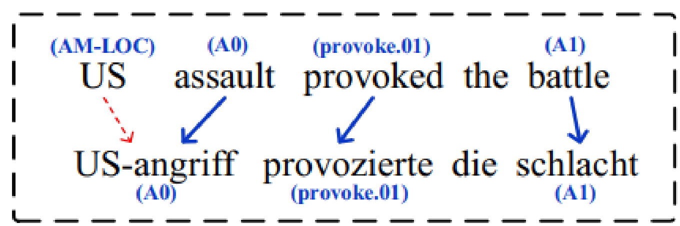
Using threshold value α to remove low confidence projections.



(a) One-to-one projection.



(b) Predicate-argument collision. Only keep predicate.



(c) Argument-argument collision. Only keep the one with higher confidence.

Semantic Role Labeling

SRL Model

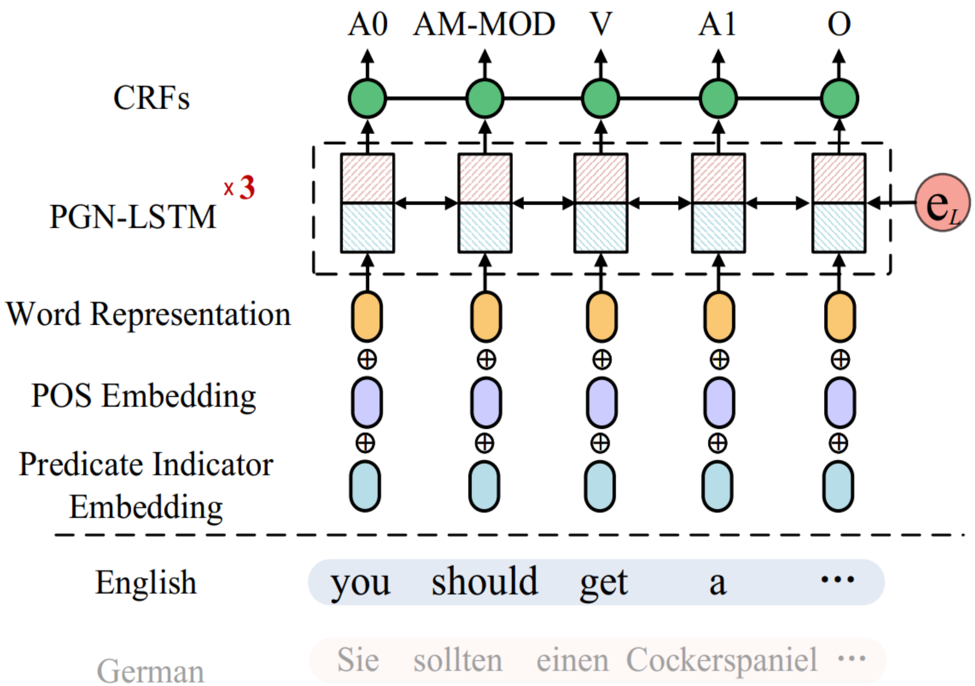
- Standard sequence labeling problem
  - Input representations:
    - word form
    - POS tag
    - predicate indicator

$$\boldsymbol{x}_i = \boldsymbol{v}_{w_i} \oplus \boldsymbol{v}_{t_i} \oplus \boldsymbol{v}_{(i==p)}$$

  - Obtain the representation via PGN:

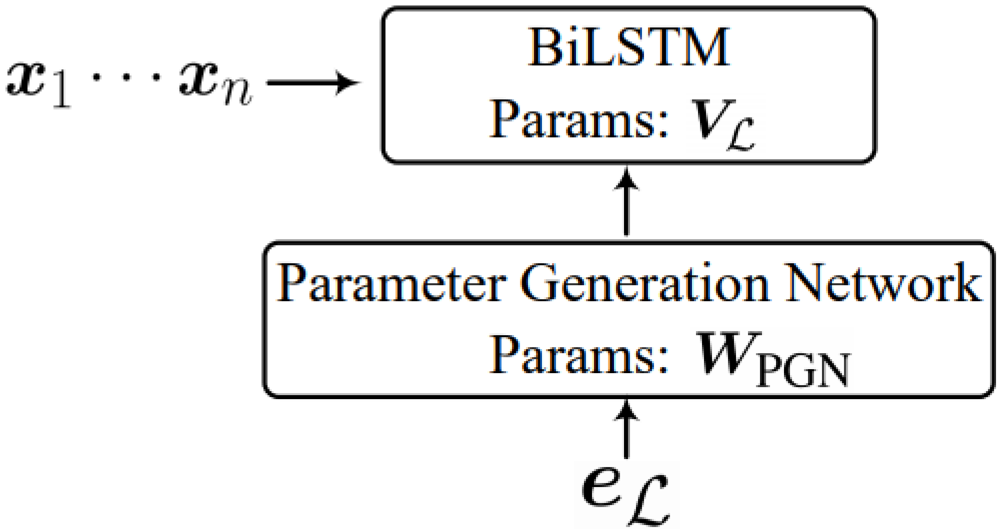
$$\begin{aligned}\boldsymbol{h}_1 \cdots \boldsymbol{h}_n &= \text{PGN-BiLSTM}(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n, \boldsymbol{e}_{\mathcal{L}}) \\ &= \text{BiLSTM}_{V_{\mathcal{L}}}(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)\end{aligned}$$

  - Decode via CRF:



Semantic Role Labeling

PGN



Semantic Role Labeling

- Experiments

➢ Universal Proposition Bank (V1.0)

Assemble the *English* based on the English EWT subset from the UDT (V1.4) and the English corpus in PB (V3.0).

Fam.	Lang.	Train	Dev	Test	Pred.	Arg.
IE.Ge	EN	10,907	1,631	1,633	41,359	100,170
	DE	14,118	799	977	23,256	58,319
IE.Ro	FR	14,554	1,596	298	26,934	44,007
	IT	12,837	489	489	26,576	56,893
	ES	28,492	3,206	1,995	81,318	177,871
	PT	7,494	938	936	19,782	41,449
Ura	FI	12,217	716	648	27,324	60,502

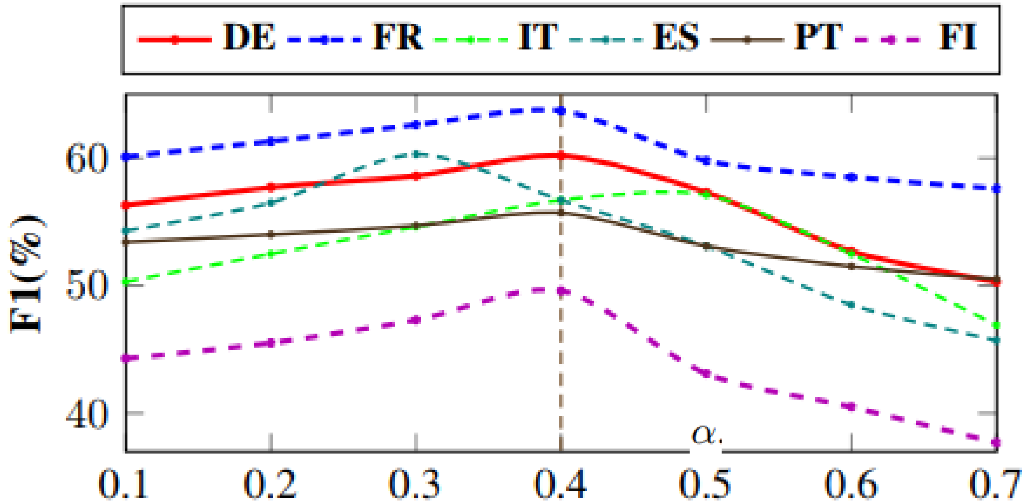
➢ Multi-lingual word form representations

(1)multi-lingual Word Embedding (Emb), via MUSE,
(2)multi-lingual ELMo, pre-trained on the blended corpus,
(3)multi-lingual BERT, officially released version (base, cased).

➢ SRL Translation

(1)Google Translation System.
(2)fastAlign for word alignment.
(3)each POS tagger trained on UDT (v1.4) for each language.

Semantic Role Labeling



Universal best projection threshold **: 0.4**

Semantic Role Labeling

- **Cross-lingual transfer from English**

  - Multilingual word representations.

    Emb < BERT < ELMo

    Surprisingly!

  - Translated target.

    TGT > SRC

  - PGN-LSTM is good.

Model	DE	FR	IT	ES	PT	FI	Avg
SRC							
Emb	42.7	51.0	42.6	40.1	43.9	30.0	41.7
BERT	43.2	53.1	44.4	41.2	44.2	31.6	43.0
ELMo	46.8	54.6	43.0	42.1	46.1	33.9	44.4
TGT							
Emb	49.4	51.3	45.5	48.4	46.9	38.7	46.7
BERT	53.0	54.3	49.1	51.3	48.8	41.1	49.6
ELMo	54.6	55.3	49.7	53.6	49.8	43.9	51.1
SRC & TGT (ELMo)							
BASIC	59.2	61.7	55.1	58.3	53.7	47.6	55.8
PGN	**65.0**	64.8	**58.7**	62.5	**56.0**	**54.5**	**60.3**
MoE	63.2	63.3	56.7	60.3	55.0	50.6	58.2
MAN-MoE	64.3	**65.3**	57.1	**62.8**	55.2	52.3	59.4

Semantic Role Labeling

➢ **Fine-grained bilingual transfer**

are all source languages useful for a target language?

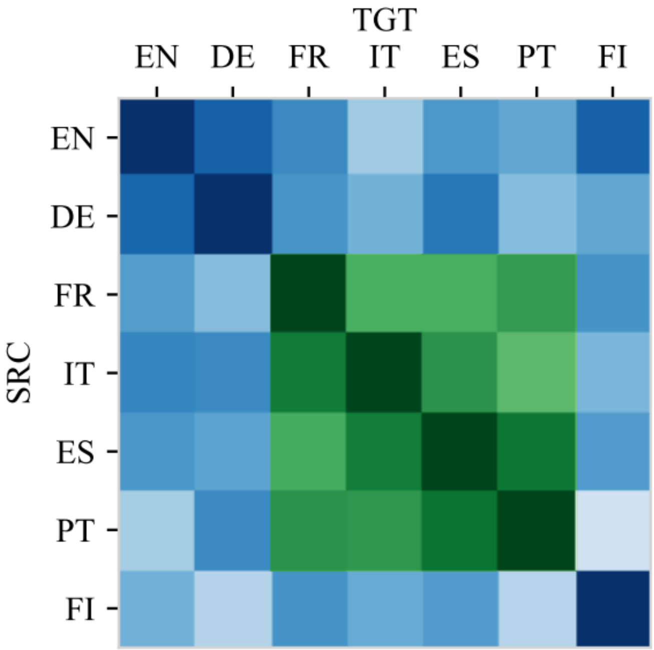
Source	EN	DE	FR	IT	ES	PT	FI
EN		**65.0**	64.8	58.7	62.5	56.0	**54.5**
DE	**63.2**		63.9	60.4	65.8	53.4	50.5
FR	60.1	53.7		63.3	63.6	62.1	51.3
IT	60.2	58.9	**65.3**		65.1	58.6	48.6
ES	60.1	57.3	64.9	**64.1**		**67.0**	50.7
PT	57.3	58.6	65.1	63.5	**67.8**		40.9
FI	50.7	52.1	64.6	53.6	60.3	51.6	
ALL	65.7	68.8	66.1	64.8	68.7	69.2	58.6

- languages belonging to one family can benefit each other greatly (i.e., EN–DE, FR–IT–ES–PT).
- multi-source transfer (i.e., All) can obtain better performances across all languages.

Semantic Role Labeling

➢ **Fine-grained bilingual transfer**

- Visualizing the language ID embeddings $e_{\mathcal{L}}$.
- Calculating the Euclidean distances in between.
- Overall tendency is highly similar to the results in **Fine-grained bilingual transfer.**
- The PGN-BiLSTM works by effectively capturing the language-aware settings.



Semantic Role Labeling

- **Bilingual transfer under all kinds of settings by each source languages**

  - languages belonging to one family can benefit each other.
  - **T**GT**&S**RC > **T**GT > **S**RC
  - Emb < BERT < ELMo
  - Language-aware encoder > Basic encoder

Target	SRC			TGT			SRC+TGT	SRC+TGT
	Emb	BERT	ELMo	Emb	BERT	ELMo	BASIC+ELMo	PGN+ELMo
Source: DE								
EN	47.32	51.62	52.82	55.04	59.20	60.48	61.05	63.21
FR	46.00	49.94	50.99	52.37	55.77	57.02	59.91	63.90
IT	40.90	43.68	45.06	48.01	51.62	52.91	57.94	60.38
ES	39.01	42.57	43.67	49.59	52.84	53.92	60.80	65.89
PT	38.25	41.73	43.07	41.44	45.76	46.94	49.14	53.40
FI	29.93	33.64	34.95	41.78	44.09	45.21	45.74	50.53
Source: FR								
EN	35.47	39.49	40.80	48.57	53.04	54.12	56.91	60.05
DE	40.01	43.86	45.24	41.33	45.16	46.54	50.53	53.69
IT	47.12	50.06	51.33	51.45	53.38	53.62	60.31	63.34
ES	40.46	44.01	45.09	50.36	53.77	54.79	58.61	63.62
PT	44.68	47.47	48.65	52.12	55.58	56.67	59.47	62.08
FI	26.44	30.92	32.07	40.71	43.97	45.05	48.76	51.31
Source: IT								
EN	37.07	39.49	40.96	47.10	51.26	52.40	54.05	60.13
DE	39.75	42.67	43.74	45.84	50.03	51.34	55.90	58.91
FR	47.39	50.08	51.28	54.45	57.29	58.78	60.03	65.30
ES	44.29	47.92	49.14	52.09	55.68	55.08	60.56	65.09
PT	42.18	46.54	47.60	49.38	53.85	54.96	57.02	58.65
FI	31.12	33.72	35.05	40.80	43.90	44.97	46.37	48.62
Source: ES								
EN	41.63	44.37	45.45	48.37	52.01	53.10	55.08	60.05
DE	36.32	39.65	40.73	44.21	47.90	49.37	51.11	57.27
FR	46.74	50.84	52.29	52.38	55.34	56.39	59.58	64.93
IT	41.39	45.42	46.82	50.10	53.01	54.01	58.83	64.09
PT	47.52	50.46	51.68	53.44	56.49	57.54	62.30	67.01
FI	29.46	32.19	33.33	39.27	42.95	44.07	47.91	50.72
Source: PT								
EN	34.83	38.09	39.27	43.16	46.09	47.50	53.12	57.30
DE	37.11	41.64	42.73	46.80	49.41	50.62	55.95	58.64
FR	42.05	46.28	47.61	49.07	52.78	54.15	58.64	65.12
IT	38.55	42.35	43.72	47.09	51.20	52.24	56.22	63.51
ES	39.58	44.01	45.02	46.57	50.61	52.06	60.84	67.81
FI	21.54	25.01	26.24	33.53	36.91	38.03	38.50	40.90
Source: FI								
EN	32.99	35.99	37.38	37.83	40.48	41.65	46.80	50.70
DE	30.98	34.59	35.70	40.75	44.41	45.84	50.68	52.12
FR	39.52	43.85	44.97	48.35	50.82	52.30	56.82	64.63
IT	33.82	36.39	37.66	41.81	46.01	47.07	52.61	53.65
ES	35.23	39.56	40.61	43.43	47.81	49.10	55.48	60.37
PT	27.93	31.90	33.30	33.84	38.21	39.43	47.75	51.61

Semantic Role Labeling

Multi-source transfer

- Similar tendencies with the single-source transfer as from English source.
- Overall performances are better then that in **Cross-lingual transfer from single English source**.

Model	EN	DE	FR	IT	ES	PT	FI	Avg
SRC								
Emb	50.3	49.2	52.4	44.9	46.7	51.0	36.4	47.3
BERT	51.8	50.6	54.0	45.3	51.3	51.8	38.1	49.0
ELMo	53.6	51.6	56.7	51.3	57.4	52.6	39.7	51.8
TGT								
Emb	56.5	51.6	55.2	47.1	50.0	53.2	40.4	50.6
BERT	59.8	55.5	57.0	52.6	54.3	56.6	44.0	54.3
ELMo	60.7	57.8	59.9	54.8	56.7	58.8	46.9	56.5
SRC & TGT (ELMo)								
BASIC	61.9	64.8	60.3	56.4	61.1	63.1	50.7	59.8
PGN	**65.7**	**68.8**	66.1	64.8	**68.7**	**69.2**	**58.6**	**66.0**
MoE	63.2	67.8	63.1	62.6	65.2	67.5	54.2	63.4
MAN-MoE	64.0	68.5	**67.2**	**65.7**	67.5	69.0	57.5	65.6

谢谢
Q/A?