A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling

YING LIN[1], SHENGQI YANG[2], VESELIN STOYANOV[3], HENG JI[1]

[1] Computer Science Department, Resselaer Polytechnic Institute

[2] Intelligent Advertising Lab, JD.com

[3] Applied Machine Learning, Facebook

MOTIVATION

- Most high-performance data-driven models rely on a large amount of labeled training data. However, a model trained on one language usually performs poorly on another language.

- Extend existing services to more languages:
  - Collect, select, and pre-process data
  - Compile guidelines for new languages
  - Train annotators to qualify for annotation tasks
  - Annotate data
  - Adjudicate annotations and assess the annotation quality and inter-annotator agreement

MOTIVATION

- Most high-performance data-driven models rely on a large amount of labeled training data. However, a model trained on one language usually performs poorly on another language.

- Extend existing services to more languages:
    - Collect, select, and pre-process data
    - Compile guidelines for new languages
    - Train annotators to qualify for annotation tasks
    - Annotate data
    - Adjudicate annotations and assess inter-annotator agreement

7,097 languages are spoken today

- Rapid and **low-cost** development of capabilities for **low-resource** languages.
    - Disaster response and recovery

TRANSFER LEARNING & MULTI-TASK LEARNING

- Leverage existing data of related languages and tasks and transfer knowledge to our target task.

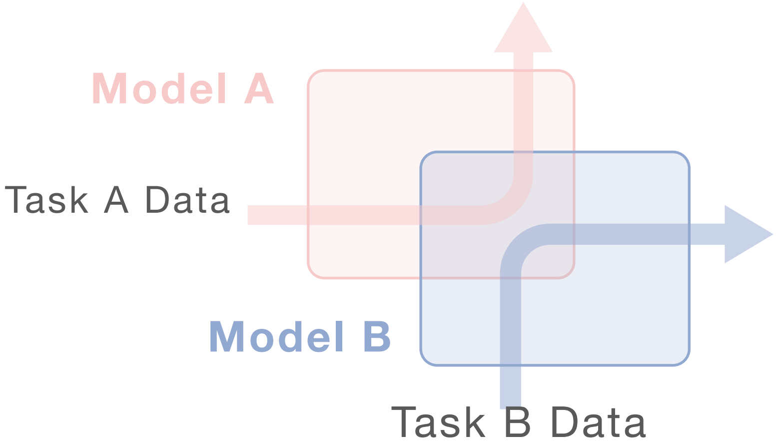
English

The Tasman Sea lies between Australia and New Zealand.

French

l'Australie est séparée de l'Asie par les mers d'Arafuraet de Timor et de la Nouvelle-Zélande par la mer de Tasman

- **Multi-task Learning** (MTL) is an effective solution for knowledge transfer across tasks.
- In the context of neural network architectures, we usually perform MTL by **sharing parameters** across models.



Parameter Sharing: When optimizing model A , we update and hence . In this way, we can partially train model B as .

SEQUENCE LABELING

- To illustrate our idea, we **take sequence labeling** as a case study.
- In the NLP context, the goal of sequence labeling is to assign a categorical label (e.g., Part-of-speech tag) to each token in a sentence.
- It underlies a range of fundamental NLP tasks, including **POS Tagging**, **Name Tagging**, and Chunking.

POS TAGGING

Koalas are largely sedentary and sleep up to 20 hours a day.

NNS VBP RB JJ CC VB IN TO CD NNS DT NN

NAME TAGGING

PER

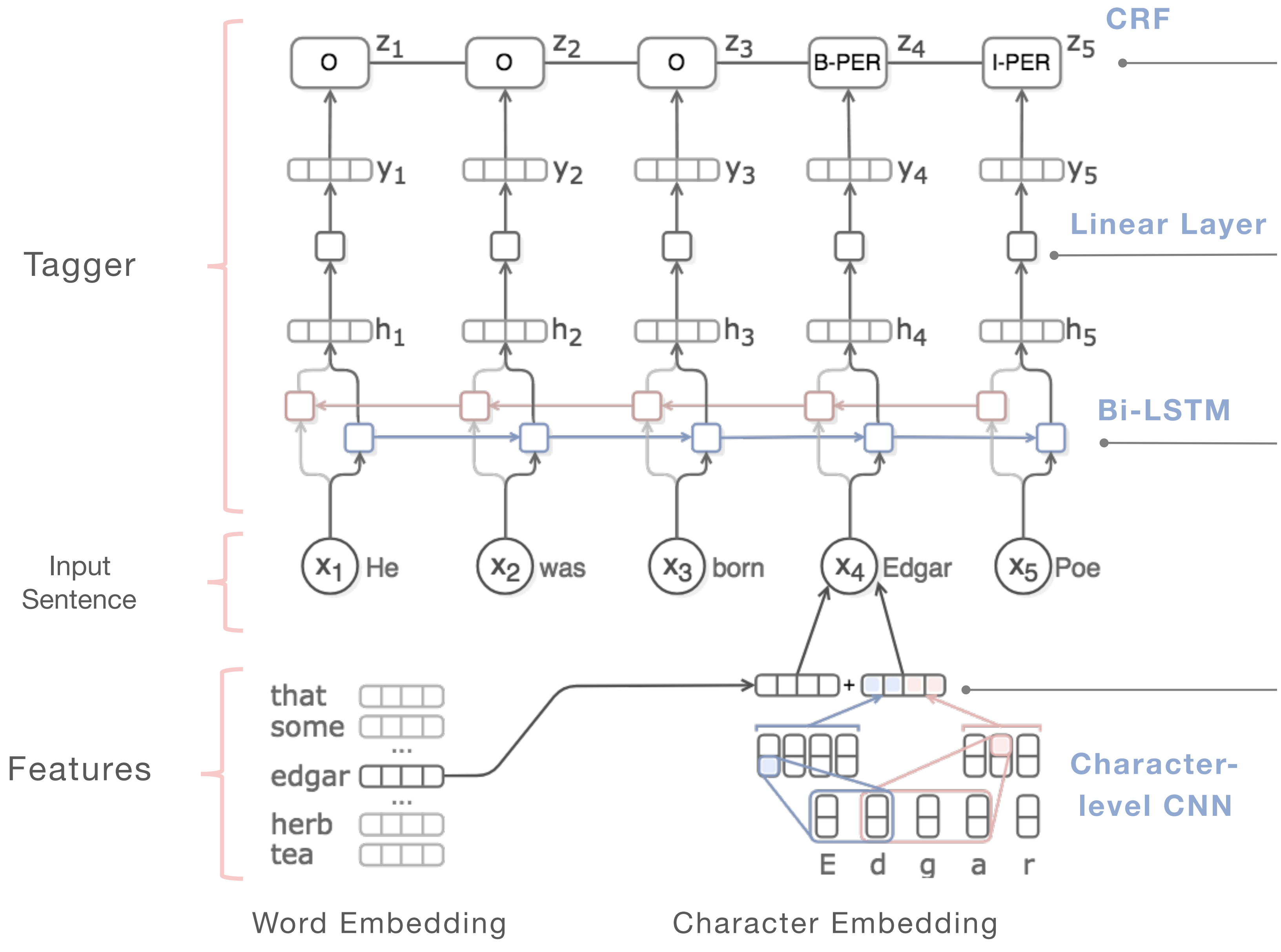
B-PER E-PER GPE GPE

Itamar Rabinovich, who as **Israel's** ambassador to **Washington** conducted unfruitful negotiations with **Syria**, told **Israel Radio** it looked like **Damascus** wated to talk rather than fight.

PER ORG GPE

- B-, I-, E-, S-: beginning of a mention, inside of a mention, the end of a mention and a single-token mention
- O: not part of any mention
- Although we only focus on sequence labeling in this work, our architecture can be adapted for many NLP tasks with slight modification.

BASE MODEL: LSTM-CRF (CHIU AND NICHOLS, 2016)



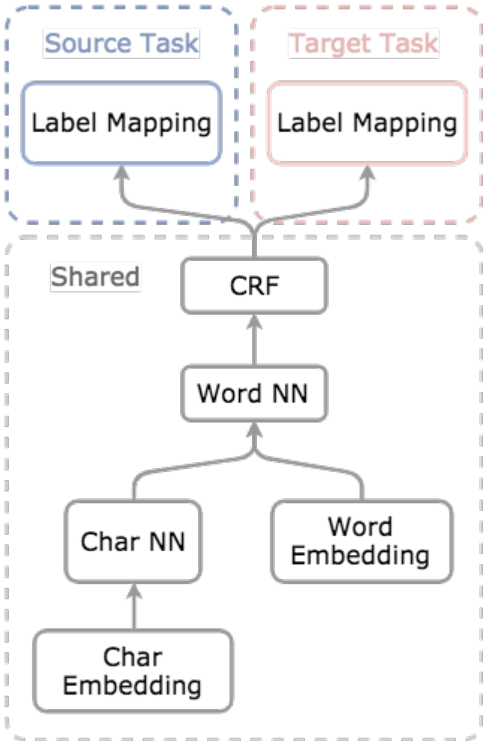
The CRF layer models the dependencies between labels.

The linear layer projects hidden states to label space.

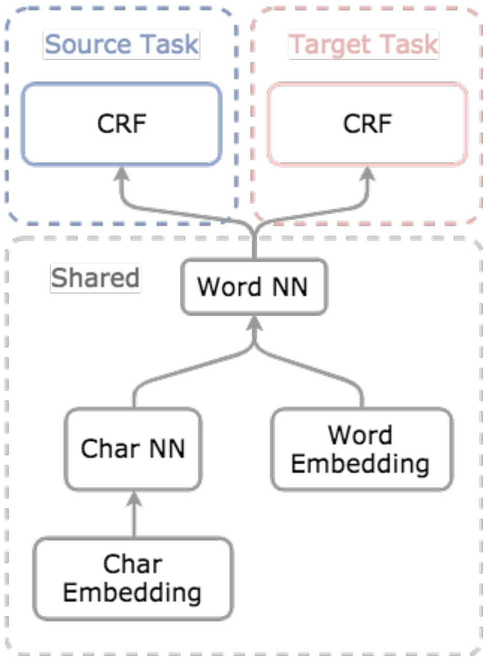
The Bidirectional LSTM (long-short term memory) processes the input sentence from both directional, encodeing each token and its context into a vector (hidden states).

Each token in the given sentence is represented as the combination of its word embedding and character feature vector.

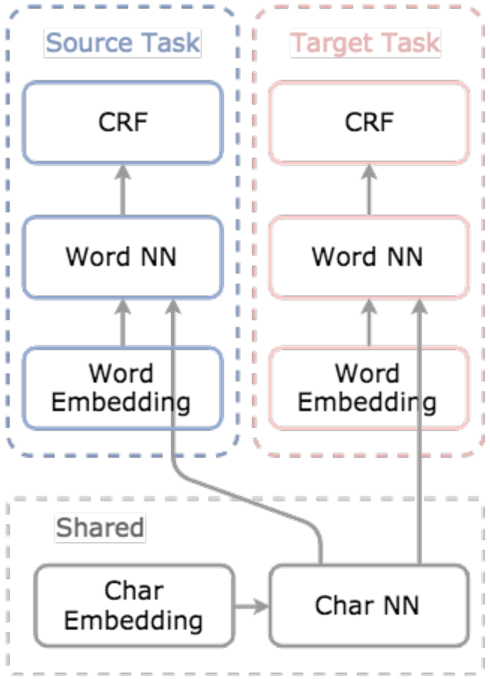
PREVIOUS TRANSFER MODELS FOR SEQUENCE LABELING



T-A: Cross-domain transfer



T-B: Cross-domain transfer With disparate label sets

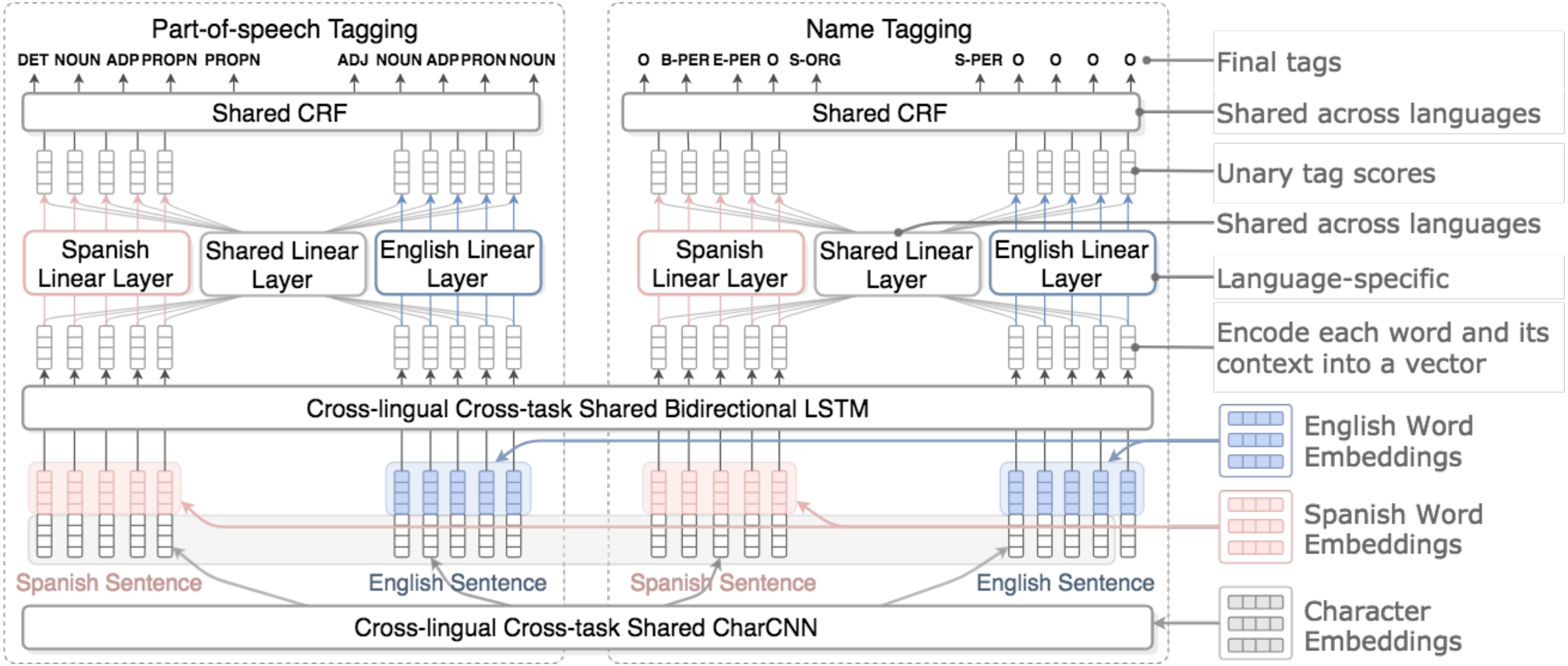


T-C: Cross-lingual Transfer

Yang et al. (2017) proposed three transfer learning architectures for different use cases.

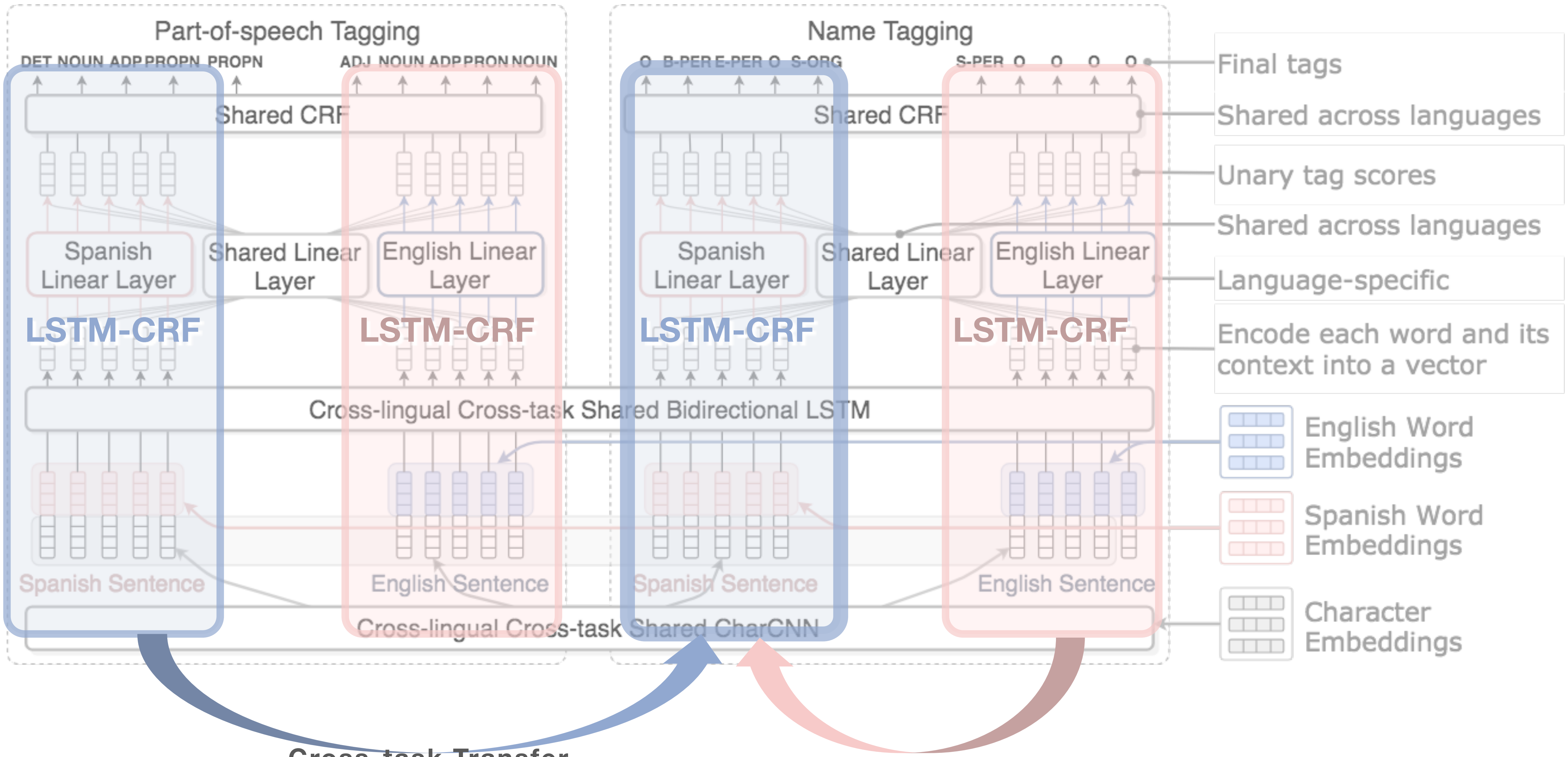
** Above figures are adapted from (Yang et al., 2017)*

OUR MODEL: MULTI-LINGUAL MULTI-TASK ARCHITECTURE



- Our model
  - combines multi-lingual transfer and multi-task transfer
  - is able to transfer knowledge from multiple sources

OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL



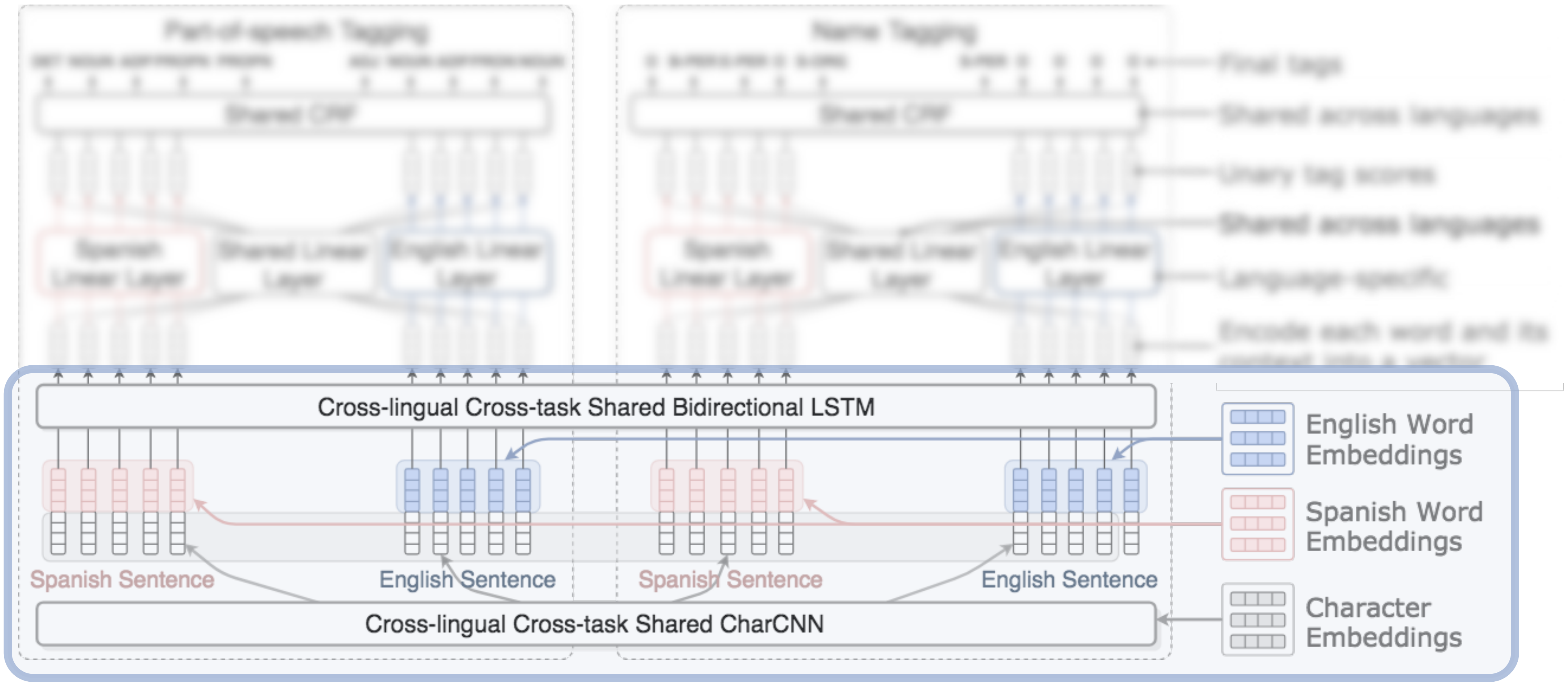
Cross-task Transfer

POS Tagging → Name Tagging

Cross-lingual Transfer

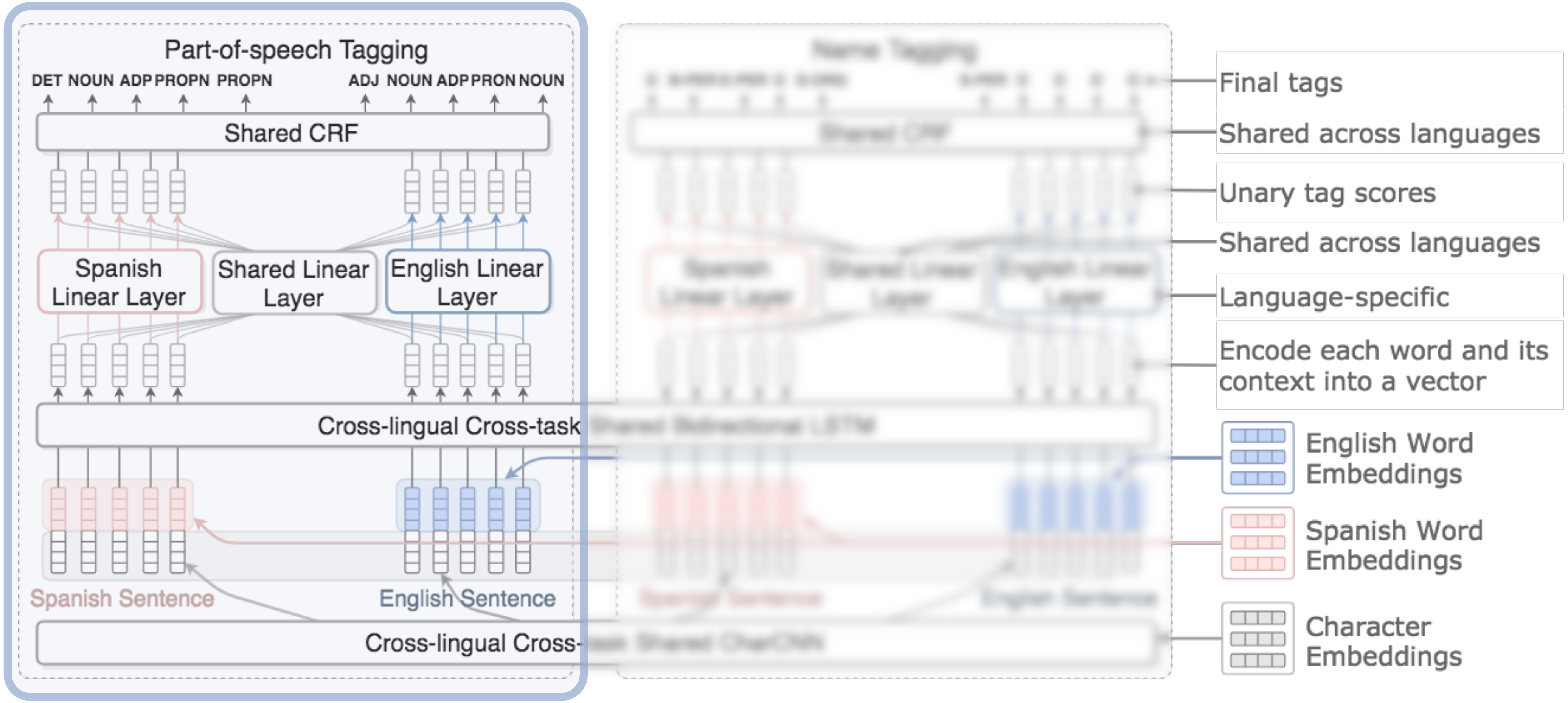
English → Spanish

OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL



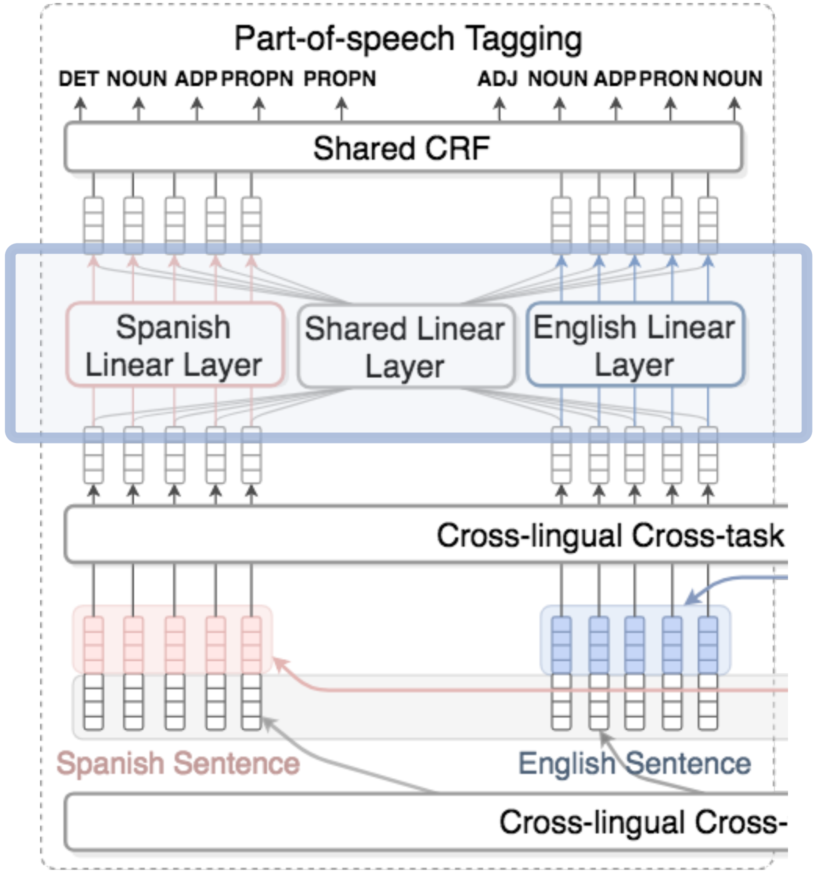
- The bidirectional LSTM, character embeddings and character-level networks serve as the basis of the architecture. This level of parameter sharing aims to provide universal word representation and feature extraction capability for all tasks and languages

OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL - CROSS-LINGUAL TRANSFER



- For the same task, most components are shared between languages.
- Although our architecture does not require aligned cross-lingual word embeddings, we also evaluate it with aligned embeddings generated using MUSE's unsupervised model (Conneau et al. 2017).

OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL - LINEAR LAYER



English: improve**ment**, develop**ment**, pay**ment**, ...

French: vrai**ment**, complète**ment**, immédiate**ment**

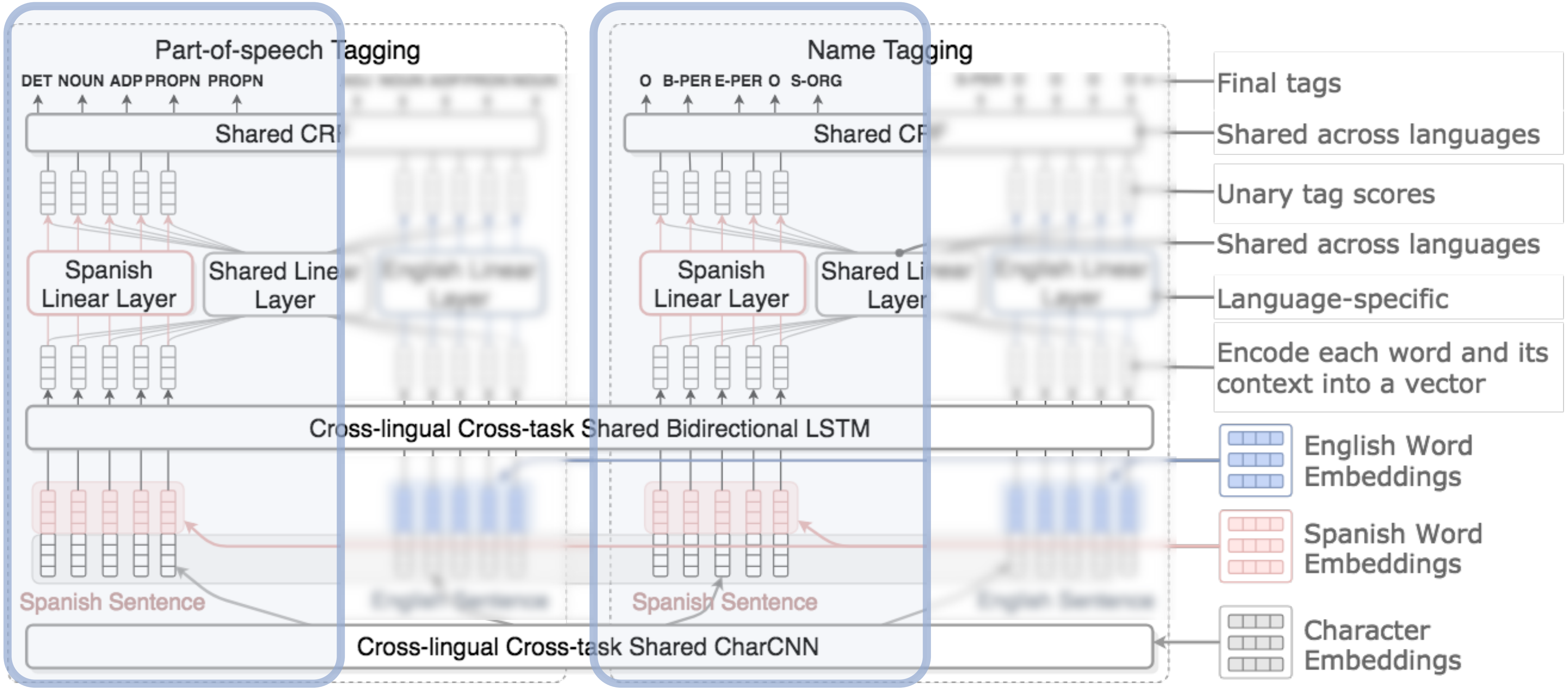
We combine the output of the shared linear layer and the output of the language-specific linear layer using

$$y = g \odot y^s + (1 - g) \odot y^u$$

where . and are optimized during training. is the LSTM hidden states. As is a square matrix, , , and have the same dimension

- We add a language-specific linear layer to allow the model to behave differently towards some features for different languages.

OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL - CROSS-TASK TRANSFER



- Linear layers and CRF layers are not shared between different tasks.
- Tasks of the same language use the same embedding matrix: mutually enhance word representations

ALTERNATING TRAINING

- To optimize multiple tasks within one model, we adopt the **alternating training** approach in (Luong et al., 2016).

$$d_1 \rightarrow d_3 \rightarrow d_2 \rightarrow d_3 \rightarrow \dots \rightarrow d_2$$

- At each training step, we sample a task with probability:

$$p(d_i) = \frac{r_i}{\sum_j r_j}$$

- In our experiments, instead of tuning mixing rate , we estimate it by:

$$r_i = \mu_i \zeta_i \sqrt{N_i}$$

where is the **task coefficient**, is the **language coefficient**, and is the **number of training examples**. (or) takes the value 1 if the task (or language) of is the same as that of the target task; Otherwise it takes the value 0.1.

EXPERIMENTS - DATA SETS

- Name Tagging
  - English: CoNLL 2003
  - Spanish and Dutch: CoNLL 2002
  - Russian: LDC2016E95 (Russian Representative Language Pack)
  - Chechen: TAC KBP 2017 10-Language EDL Pilot Evaluation Source Corpus
- Part-of-speech Tagging: CoNLL 2017 (Universal Dependencies)

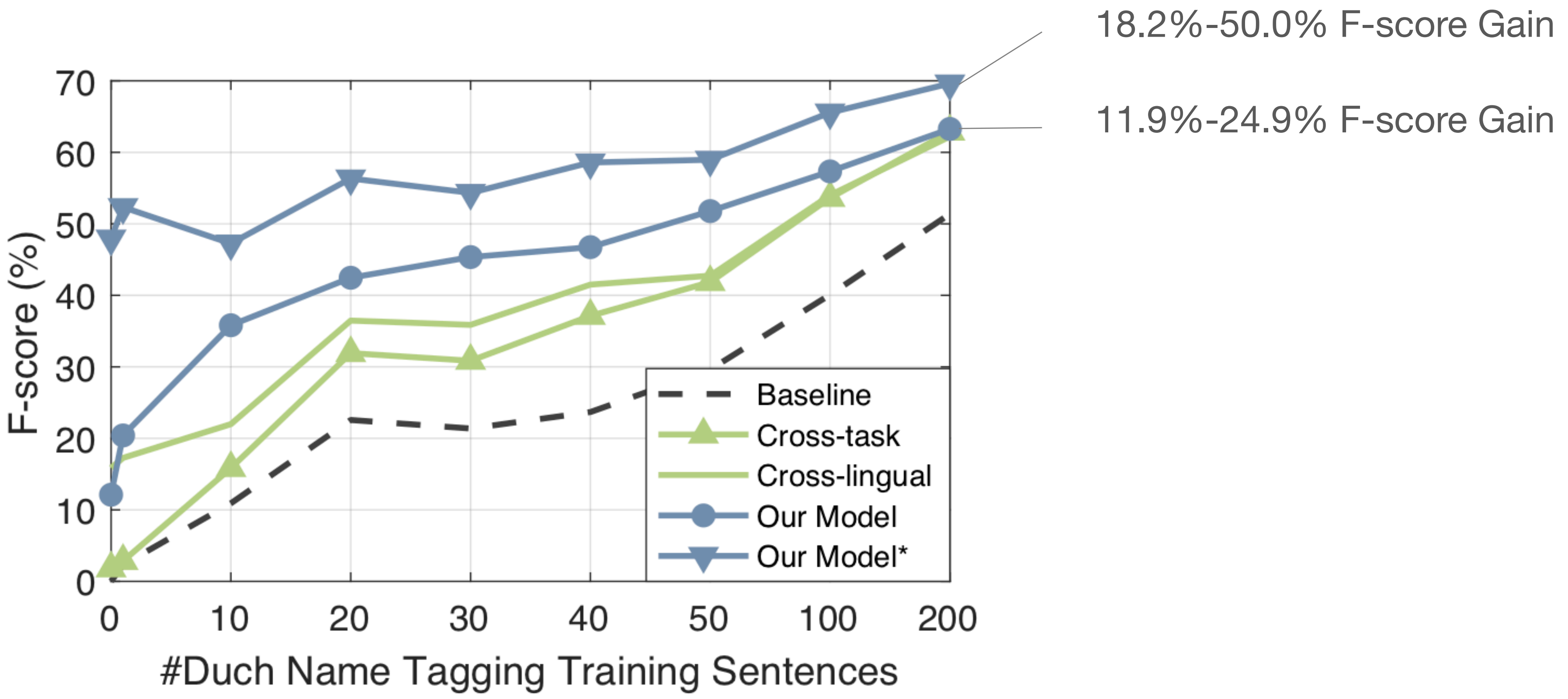
EXPERIMENTS - SETUP

- 50-dimensional pre-trained word embeddings
  - English, Spanish and Dutch: Wikipedia
  - Russian: LDC2016E95
  - Chechen: TAC KBP 2017 10-Language EDL Pilot Evaluation Source Corpus
- Cross-lingual word embedding: we aligned mono-lingual pre-trained word embeddings with MUSE (https://github.com/facebookresearch/MUSE).
- 50-dimensional randomly initialized character embeddings
- Optimization: SGD with momentum (), gradient clipping (threshold: 5.0) and exponential learning rate decay.

CharCNN Filter Number	20
Highway Layer Number	2
Highway Activation Function	SeLU
LSTM Hidden State Size	171
LSTM Dropout Rate	0.6
Learning Rate	0.02
Batch Size	19

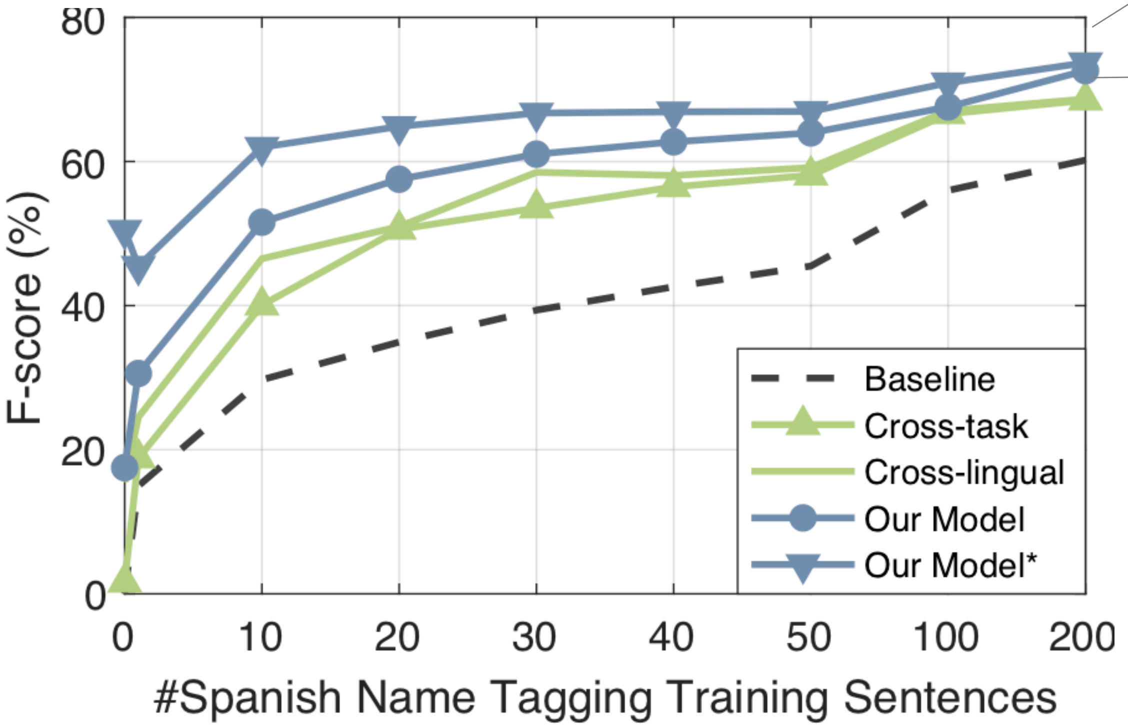
EXPERIMENTS - COMPARISON OF DIFFERENT MODELS

- Target task: Dutch Name Tagging
- Auxiliary task: Dutch POS Tagging, English Name Tagging, English POS Tagging



EXPERIMENTS - COMPARISON OF DIFFERENT MODELS

- Target task: Spanish Name Tagging
- Auxiliary task: Spanish POS Tagging, English Name Tagging, English POS Tagging

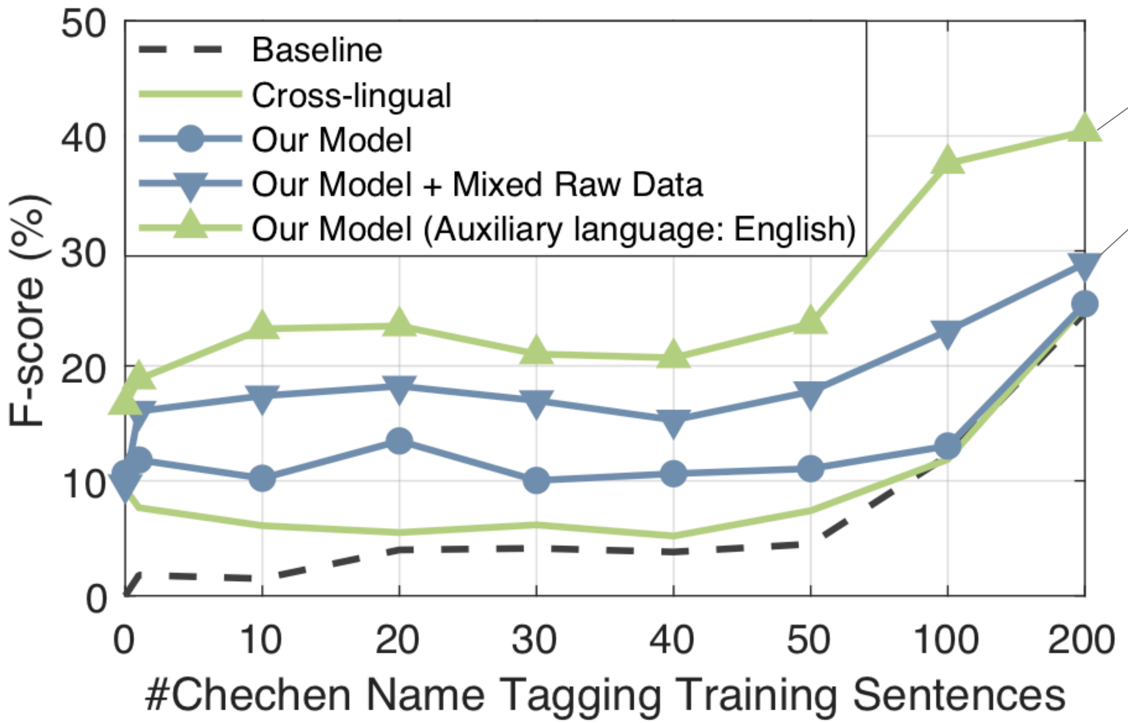


13.5%-50.5% F-score Gain

11.6%-22.6% F-score Gain

EXPERIMENTS - COMPARISON OF DIFFERENT MODELS

- Target task: Chechen Name Tagging
- Auxiliary task: Russian POS Tagging + Name Tagging or English POS Tagging + Name Tagging



15.8%-25.4% F-score Gain

4.3%-15.9% F-score Gain

All training data:

Baseline: 78.9%

Our Model : 82.3%

EXPERIMENTS - COMPARISON WITH STATE-OF-THE-ART MODELS

Language	Model	F-score
Dutch	Glilick et al. (2016)	82.84
	Lample et al. (2016)	81.74
	Yang et al. (2017)	85.19
	Baseline	85.14
	Cross-task	85.69
	Cross-lingual	85.71
	Our Model	**86.55**
Spanish	Glilick et al. (2016)	82.95
	Lample et al. (2016)	85.75
	Yang et al. (2017)	85.77
	Baseline	85.44
	Cross-task	85.37
	Cross-lingual	85.02
	Our Model	**85.88**

- We also compared our model with state-of-the-art models with all training data.

EXPERIMENTS - COMPARISON WITH STATE-OF-THE-ART MODELS

#1 [DUTCH]: *If a Palestinian State is, however, the first thing the Palestinians will do.*

⋆ [B] Als er een Palestijnse staat komt, is dat echter het eerste wat de Palestijnen zullen doen

⋆ [A] Als er een [S-MISC Palestijnse] staat komt, is dat echter het eerste wat de [S-MISC Palestijnen] zullen doen

#2 [DUTCH]: *That also frustrates the Muscovites, who still live in the proud capital of Russia but can not look at the soaps that the stupid farmers can see on the outside.*

⋆ [B] Ook dat frustreert de Moskovieten , die toch in de fiere hoofdstad van Rusland wonen maar niet naar de soaps kunnen kijken die de domme boeren op de buiten wel kunnen zien

⋆ [A] Ook dat frustreert de [S-MISC Moskovieten] , die toch in de fiere hoofdstad van [S-LOC Rusland] wonen maar niet naar de soaps kunnen kijken die de domme boeren op de buiten wel kunnen zien

#3 [DUTCH]: *And the PMS centers are merging with the centers for school supervision, the MSTs.*

⋆ [B] En smelten de PMS-centra samen met de centra voor schooltoezicht, de MST's .

⋆ [A] En smelten de [S-MISC PMS-centra] samen met de centra voor schooltoezicht, de [S-MISC MST's] .

#4 [SPANISH]: *The trade union section of CC.OO. in the Department of Justice has today denounced more attacks of students to educators in centers dependent on this department ...*

⋆ [B] La [B-ORG sección] [I-ORG sindical] [I-ORG de] [S-ORG CC.OO.] en el [B-ORG Departamento] [I-ORG de] [E-ORG Justicia] ha denunciado hoy ms agresiones de alumnos a educadores en centros dependientes de esta [S-ORG consellería] ...

⋆ [A] La sección sindical de [S-ORG CC.OO.] en el [B-ORG Departamento] [I-ORG de] [E-ORG Justicia] ha denunciado hoy ms agresiones de alumnos a educadores en centros dependientes de esta consellería ...

#5 [SPANISH]: ... *and the Single Trade Union Confederation of Peasant Workers of Bolivia, agreed upon when the state of siege was ended last month.*

⋆ [B] ... y la [B-ORG Confederación] [I-ORG Sindical] [I-ORG Unica] [I-ORG de] [E-ORG Trabajadores] Campesinos de [S-ORG Bolivia] , pactadas cuando se dio fin al estado de sitio, el mes pasado .

⋆ [A] .. y la [B-ORG Confederación] [I-ORG Sindical] [I-ORG Unica] [I-ORG de] [I-ORG Trabajadores] [I-ORG Campesinos] [I-ORG de] [E-ORG Bolivia] , pactadas cuando se dio fin al estado de sitio, el mes pasado .

Baseline

Our Model

Incorrect

Correct

EXPERIMENTS - CROSS-TASK TRANSFER VS CROSS-LINGUAL TRANSFER

[DUTCH] … *Ingeborg Marx is her name, a formidable heavy weight to high above her head!*

⋆ [B] … Zag ik zelfs onlangs niet dat een lief, mooi vrouwtje, Ingeborg Marx is haar naam, een formidabel zwaar gewicht tot hoog boven haar hoofd stak!

⋆ [CROSS-TASK] … Zag ik zelfs onlangs niet dat een lief, mooi vrouwtje, [B-PER Ingeborg] [S-PER Marx] is haar naam, een formidabel zwaar gewicht tot hoog boven haar hoofd stak!

⋆ [CROSS-LINGUAL] … Zag ik zelfs onlangs niet dat een lief, mooi vrouwtje, [B-PER Ingeborg] [E-PER Marx] is haar naam, een formidabel zwaar gewicht tot hoog boven haar hoofd stak!

- With 100 Dutch training sentences:
  - The baseline model misses the name "Ingeborg Marx".
  - The cross-task transfer model finds the name but assigns a wrong tag to "Marx".
  - The cross-lingual transfer model correctly identifies the whole name.
- The task-specific knowledge that B-PER → S-PER is an invalid transition will not be learned in the POS Tagging model.
- The cross-lingual transfer model transfers such knowledge through the shared CRF layer.

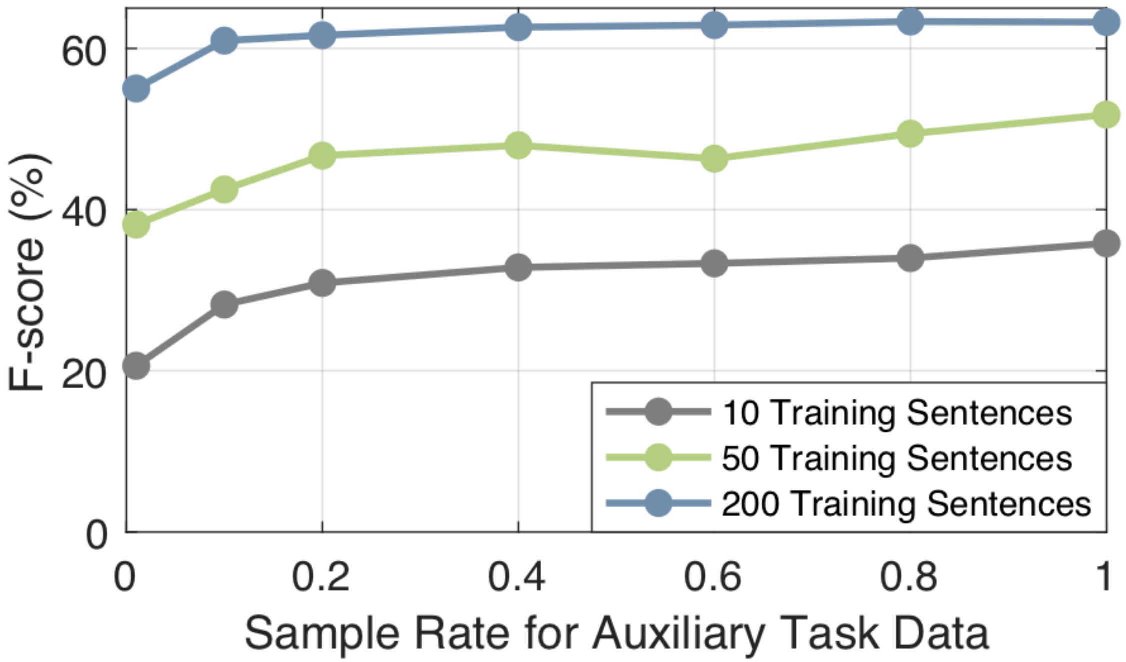
EXPERIMENTS - ABLATION STUDIES

Model	0	10	100	200	All
Basic	2.06	20.03	47.98	51.52	77.63
+C	1.69	24.22	48.53	56.26	83.38
+CL	9.62	25.97	49.54	56.29	83.37
+CLS	3.21	25.43	50.67	56.34	84.02
+CLSH	7.70	30.48	53.73	58.09	84.68
+CLSHD	12.12	35.82	57.33	63.27	86.00

C: Character embedding; L: Shared LSTM; S: Language-specific
H: Highway Networks; D: Dropout

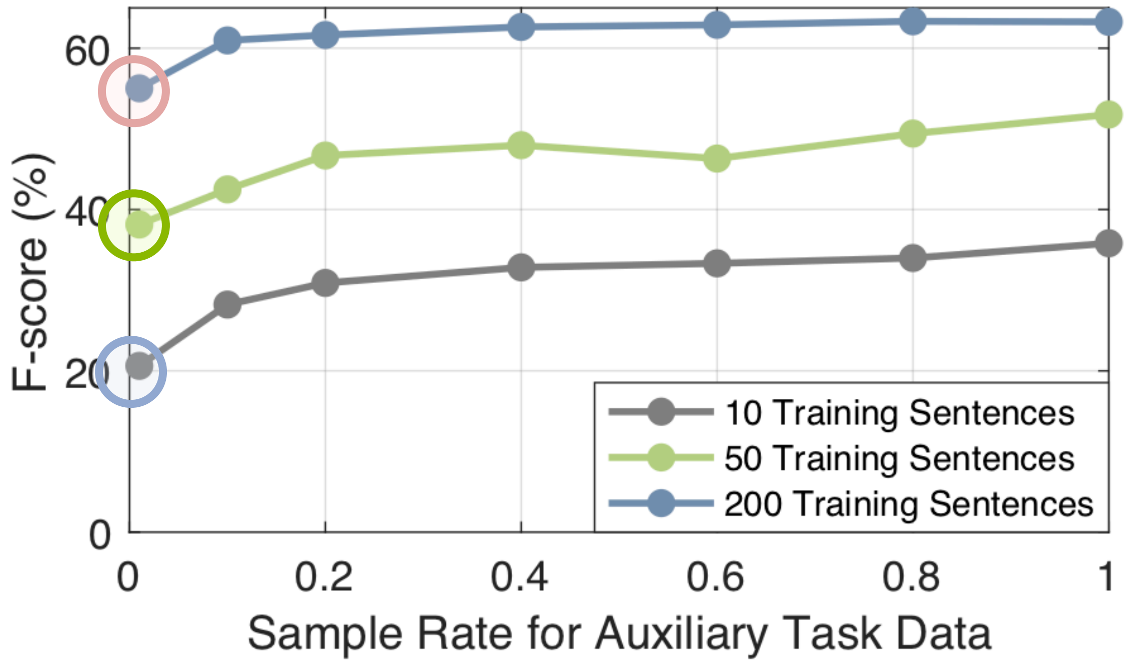
- Generally, all components improve the performance.
- Sharing the LSTM layer slightly hurts the performance in the "high-resource" setting.
- Language-specific Layer can impair the performance in extreme low-resource settings because this layer is trained only on the target task data.

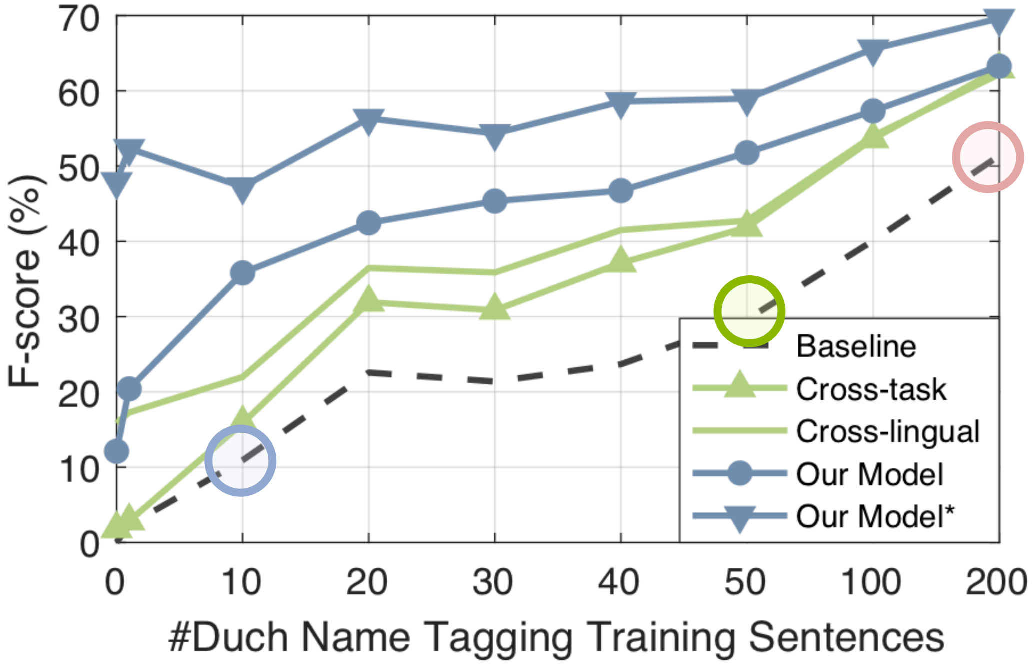
EXPERIMENTS - EFFECT OF THE AMOUNT OF AUXILIARY TASK DATA



- Does our model heavily rely on the amount of auxiliary task data?
  - The performance goes up when we increase the sample rate from 0 to 0.2 for auxiliary task data.
  - However, we do not observe substantial improvement when we further increase the sample rate.
- Using only 1% auxiliary data, our model already obtains 3.7%-9.7% absolute F-score gains.

EXPERIMENTS - EFFECT OF THE AMOUNT OF AUXILIARY TASK DATA





- Does our model heavily rely on the amount of auxiliary task data?
  - The performance goes up when we increase the sample rate from 0 to 0.2 for auxiliary task data.
  - However, we do not observe substantial improvement when we further increase the sample rate.
- Using only 1% auxiliary data, our model already obtains 3.7%-9.7% absolute F-score gains.

REFERENCES

- Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. TACL, 4:357–370
- Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herve J ´ egou. 2017. ´ Word translation without parallel data. arXiv preprint arXiv:1710.04087
- Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In NAACL HLT
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In NAACL HLT
- Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In ICLR

Thank you ☺️

A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling

YING LIN, SHENGQI YANG, VESELIN STOYANOV, HENG JI