Multilingual Complex Named Entity Recognition via Data Augmentation

Oreen Yousuf

UPPSALA
UNIVERSITET

Background – (Complex) NER

Named Entity Recognition (NER)

- Foundational task in LT – Information Extraction, spell-checkers, dialogue systems, search, etc. heavily rely on NER

Complex, Ambigious, Emerging Entities:

- Low-context settings
- Any linguistic form
- Complex Entites could be complex noun phrases from creative works (e.g. movies, literature) – ”Eternal Sunshine of the Spotless Mind”, an infinitive – ”To Sleep with Anger”, or even gerunds – ”The Coming of Sin”, etc.
- Ambiguous Entities – difficult to recognize due to context, e.g., ”Frozen” (movie), ”Among Us” (video game), ”It’s Always Sunny in Philadelphia” (TV show), etc.
- Emerging Entites – perpetually created

Background – Data Augmentation

Process of altering original data to create new data instances. Consolidate new + synthetic data to bolster resources that will be used

Much recent attention on sentence-level and sentence-pair NLP tasks, e.g., MT, NLI, text classification

- Manipulation: Word Replacement, Random Deletion, Swap Word Position
- Artificial Creation: Generative models (variational auto encoders) or Back-Translation

Data augmentation for NER is comparatively less studied:

- Predictions at token level determine if it is a mention and which entity type the mention has
- Transforming tokens may change labels

Augmentation Techniques

Entitiy Types:

- PER (Person)
- LOC (Location)
- GRP (Group)
- CORP (Corporation)
- PROD (Product)
- CW (Creative Work)

Original	all	songs	written	by	m.o.d.	,	unless	otherwise	stated	
label	O	O	O	O	B-GRP	O	O	O	O	
LwTR	**a**	songs	written	by	**england**	,	**drowned**	otherwise	**contested**	
label	O	O	O	O	B-GRP	O	O	O	O	
SR	all	**song**	**written—**	**was**	m.o.d.	.	**if**	otherwise	**said**	
label	O	O	O	O	B-GRP	O	O	O	O	
MR	all	songs	written	by	**es**	**tunis**	.	unless	otherwise	stated
label	O	O	O	O	B-GRP	I-GRP	O	O	O	O
SiS	all	songs	written	by	m.o.d.	**otherwise**	**unless**	**stated**	**,**	
label	O	O	O	O	B-GRP	O	O	O	O	
All	**songs**	**was**	**a**	**written—**	**england**	**otherwised**	**.**	**contested**	**drowned**	
label	O	O	O	O	B-GRP	O	O	O	O	

- Label-wise Token Replacement – randomly replace token with another token of the same label type
- Synonym Replacement – randomly replace token with a synonym of itself via fastText word embedding
- Mention Replacement – randomly replace an entire named entity with another entity of the same label type
- Shuffle Within Segments – divide sentence into segments of the same label, randomly shuffle segments
- * All – perform all 4 techniques on a single dataset
- * Combined – with 4 copies of the original dataset, perform the 4 techniques and append each to the original, yielding a final dataset 5x as long as the original

UPPSALA UNIVERSITET

네덜란드의	축구	선수	사리	판	페이넨달	.
O	O	O	B-PER	I-PER	I-PER	O
네덜란드의	은	선수	사리	**가와쿠보**	페이넨달	.
O	O	O	B-PER	I-PER	I-PER	O

Korean LwTR

它	以	其	大	西	洋	鮭	钓	鱼	。
O	O	O	B-PROD	I-PROD	I-PROD	I-PROD	O	O	O
它	以	**而**	**小**	**南**	**澄**	**紅鮭**	**钓上**	鱼	。
O	O	O	B-PROD	I-PROD	I-PROD	I-PROD	O	O	O

Chinese SR

Bangla MR

তিনি	মার্কো	মারুলি	লেখকের	সহযোগী	ছিলেন	বলেও	জানা	যায়।
O	B-PER	I-PER	O	O	O	O	O	O
তিনি	**ব্রুস**	**মাসলিন**	লেখকের	সহযোগী	ছিলেন	বলেও	জানা	যায়।
O	B-PER	I-PER	O	O	O	O	O	O

Spanish SiS

ese	mismo	año	también	se	publicó	amused	to	death	,	el	nuevo	trabajo	en	solitario	de	roger	waters	.
O	O	O	O	O	O	B-CW	I-CW	I-CW	O	O	O	O	O	O	O	B-PER	I-PER	O
ese	mismo	año	también	se	publicó	**to**	**amused**	**death**	**de**	**trabajo**	**solitario**	**el**	**nuevo**	,	**en**	roger	waters	.
O	O	O	O	O	O	B-CW	I-CW	I-CW	O	O	O	O	O	O	O	B-PER	I-PER	O

UPPSALA UNIVERSITET

Environmental Setup

SemEval 2022 Task 11 Dataset (Wikipedia & Microsoft Orcas):

	Train	Dev.	Test
Each Language	13,770	800	1,530

Entitiy Types:

- PER (Person)
- LOC (Location)
- GRP (Group)
- CORP (Corporation)
- PROD (Product)
- CW (Creative Work)

XLM-RoBERTa:

- Multilingual Transformer-based masked language model
  - Excludes linear chain CRF
  - Masked language modeling (MLM); randomly masks % of input sentence's words before feeding into the model for predictions on masked words
- Prominent model used for NER
- Trained on 2.5TB of filtered CommonCrawl data

fastText word embedding (trained on Wikipedia)

Environmental Setup

Group 1: Korean & Turkish

- Synthetic languages; high use of inflection or agglutination to specify syntactic relationship between a sentence's words. Very flexible sentence structure

Group 2: Bangla, Hindi & Farsi

- Common SOV order, Indo-Iranian, BN/HI especially geographically close

Group 3: German, Russian, Dutch, Spanish & English

- Geographically close European languages with shared SVO* grammar, 3 of which are Germanic

Chinese used as a test lanugage due to its uniqueness in relation to the 10 other languages

*German and Dutch are SVO in conventional typology and SOV in generative grammar

Initial Results

Epochs = 2, lr = 0.0001, batch size = 64, dropout = 0.1											
	KO	TR	*ZH*	**BN**	HI	FA	**DE**	RU	NL	ES	EN
original	0.646	0.668	0.638	0.528	0.548	0.593	0.716	0.631	0.683	0.654	0.696
LwTR	0.495	0.556	0.390	0.390	0.381	0.487	0.551	0.418	0.572	0.531	0.540
MR	**0.617**	**0.675**	**0.638**	0.426	**0.548**	**0.642**	**0.723**	**0.632**	**0.700**	**0.709**	**0.709**
SR	0.585	0.231	0.431	**0.498**	0.407	0.582	0.475	0.507	0.662	0.602	0.629
SiS	0.537	0.564	0.365	0.379	0.404	0.543	0.557	0.422	0.580	0.516	0.532
All	0.384	0.442	0.184	0.178	0.415	0.351	0.424	0.301	0.465	0.396	0.414

Table 3: Evaluation results of micro F1 scores across all six named entity labels. All models are trained on a single language indicated by their corresponding language code. Bold language codes indicate beginning of language groupings (1. **Korean**/Turkish, 2. **Bangla**/Hindi/Farsi, 3. **German**, Russian, Dutch, Spanish, English). Chinese (*ZH*) is stand-alone. Bold indicates best performing method for a language in the Table.

Mention Replacement (MR) best performing augmentation method for 10/11 languages

- In the 10/11 languages, MR +10.67% > 2nd best method
- Better than original data: 7/11
- As good as original: 2/11
- Worse than original: 2/11 (Korean and Bangla)

Tally of worst augmentation: LwTR (5), SiS (4), SR (2)

Most gained from MR: Spanish (ES): +5.4%, Farsi (FA): +4.9%

Best augmentation for Bangal (BN) < Original data

- Surprisingly best performing augmentation was Synonym Repalcement (SR); word embeddings for low-resource languages aren't always reliable

'All' is by far the worst approach

UPPSALA UNIVERSITET

Combined Augmentation & Cross-Lingual Results

Epochs = 2, lr = 0.0001, batch size = 64, dropout = 0.1											
Language	**KO**	TR	*ZH*	**BN**	HI	FA	**DE**	RU	NL	ES	EN
Original	0.646	0.668	0.638	0.528	0.548	0.593	0.716	0.631	0.683	0.654	0.696
Combined	**0.735**	**0.745**	0.589	0.428	**0.643**	0.420	0.726	0.712	0.715	0.717	0.724
Group Train	0.679	0.683	N/A	**0.614**	0.606	**0.626**	**0.739**	**0.753**	**0.734**	**0.747**	**0.752**

Table 4: Combined indicates the model training on a single language's original data plus each of its augmented datasets. KO's Combined score of 0.735 is the micro F1 score across 6 NER labels when trained on the Original KO data + LwTR KO data + MR KO data + SR KO data + SiS KO data. Group Train indicates the result from training on original training data from each of a group's languages. KO's Group Train score of 0.679 is the micro F1 score across 6 NER labels when trained on KO and TR's original data, as KO and TR are in the same group - this group training logic of course applies for the BN/HI/FA, and DE/RU/NL/ES/EN groups. Chinese (ZH) is excluded as it was not placed in a language group. Bold indicates best performing method for a language in the Table.

While Groups 1 & 3 consistently prefer one method over the other, their worse performing method still outperformed both their previous single best method and original data in all languages.

- Suggests Group 1 benefits most with larger amounts of its own data, while Group 3 can find improvements with less data via cross-lingual training

Chinese performed worse than previous MR

Bangla now surpasses original data score with cross-lingual training from Group 2 languages HI & FA

No clear consistency in Group 2; comparatively less insight than Group 1 & 3.

- Hindi: Cross-lingual training still outperformed all scores from previous experiments

Script Based Results

Epochs = 2, lr = 0.0001, batch size = 64, dropout = 0.1					
Latin Script Lang.	Turkish	German	Dutch	Spanish	English
individual original	0.668	0.716	0.683	0.654	0.696
all latin original	0.760	0.740	0.761	0.751	0.750
euro latin original	N/A	0.756	0.773	0.766	0.763
all latin mr	**0.780**	0.778	0.770	0.770	0.771
euro latin mr	N/A	**0.797**	**0.782**	**0.801**	**0.797**

Table 5: We reiterate the languages' original results from the previous page in the first row of results for ease of comparison. The 'all latin original' and 'all latin mr' were trained on all 5 languages' original and MR datasets, respectively. The 'euro latin original' and 'euro latin mr' were trained on just the European languages' original and MR datasets, respectively (i.e., excluding Turkish). Bold indicates best performing method for a language in the Table.

Performances increases as you move down the rows for each language

Spanish reaches best performance among all experiments so far when cross-lingually trained with Latin orthography-based European languages

Second pattern:

- ”all latin mr” > ”all latin original” & ”euro latin mr” > ”euro latin original” for European languages

Similarly, Turkish MR evaluations > original

Euro Script Based Results had less *compiled* data than the previous Combined model, implying larger amounts of augmented data can be outperformed

UPPSALA UNIVERSITET

Spanish Label Analysis

Juxtapose original, MR, and Best (”euro latin mr”) label results

Best outperforms original across the board, while MR does so with the exception of the CORP (Corporation) label

MR performs worse than original for CORP by ~2.6% in all metrics

- Contrasting Chinese breakdown which showcases instances of nonhegemonic metric outperformances (i.e., worse MR recall doesn’t necessarily mean worse MR precision)

Entity	Metric	Original	MR	Best
PROD	P	0.524	**0.644**	**0.745**
	R	0.457	**0.645**	**0.719**
	F1	0.488	**0.644**	**0.732**
GRP	P	0.615	**0.693**	**0.797**
	R	0.551	**0.649**	**0.758**
	F1	0.581	**0.670**	**0.777**
CORP	P	0.682	0.657	**0.771**
	R	0.661	0.633	**0.758**
	F1	0.671	0.644	**0.765**
CW	P	0.547	**0.589**	**0.715**
	R	0.471	**0.557**	**0.687**
	F1	0.506	**0.573**	**0.700**
PER	P	0.784	**0.808**	**0.890**
	R	0.822	**0.860**	**0.911**
	F1	0.802	**0.833**	**0.900**
LOC	P	0.749	**0.779**	**0.848**
	R	0.735	**0.771**	**0.847**
	F1	0.742	**0.775**	**0.848**
Micro Avg.	P	0.672	**0.713**	**0.807**
	R	0.636	**0.706**	**0.795**
	F1	0.654	**0.709**	**0.801**
MD	P	0.762	**0.837**	**0.878**
	R	0.721	**0.829**	**0.865**
	F1	0.741	**0.833**	**0.872**

Table 6: Breakdown of individual Spanish (ES) label results comparing the original training data, its best performance under a single data augmentation method - Mention Replacement (MR) - and its best overall data augmentation method - euro latin mr - as seen in Table 5. Bold indicates if a metric performs better than or as good as the original

UPPSALA UNIVERSITET

Chinese Label Analysis

While Combined micro F1 5% < original:

Combined GRP (Group) was substantially better; +16% recall gain

GRP is worst performing entity in all experiments

Avg. Precision-recall difference for other entities is 2.52%, while GRP's is 27.6%

Combined's larger datasize could be deciding factor in recall performance

Precision for Chinese GRP can perform well with a leading augmentation method, recall plummets without bolstering data

Entity	Metric	Original	LwTR	MR	SR	SiS	All	Combined
PROD	P	0.574	0.305	**0.586**	0.368	0.351	0.171	0.552
	R	0.534	0.282	0.530	0.291	0.313	0.134	0.514
	F1	0.553	0.293	**0.557**	0.325	0.330	0.150	0.533
GRP	P	0.493	0.191	**0.530**	0.312	0.262	0.176	**0.513**
	R	0.217	0.035	**0.225**	0.0162	0.0733	0.00493	**0.377**
	F1	0.301	0.0591	**0.316**	0.0308	0.115	0.00958	**0.435**
CORP	P	0.644	0.370	**0.644**	0.433	0.338	0.164	0.572
	R	0.619	0.349	0.604	0.381	0.339	0.149	0.557
	F1	0.631	0.359	0.623	0.405	0.338	0.156	0.564
CW	P	0.525	0.239	**0.531**	0.303	0.297	0.113	0.513
	R	0.542	0.219	0.525	0.255	0.271	0.0703	0.483
	F1	0.533	0.228	0.528	0.277	0.284	0.0868	0.498
PER	P	0.753	0.611	**0.774**	0.590	0.431	0.297	0.696
	R	0.769	0.610	**0.777**	0.642	0.448	0.291	0.708
	F1	0.761	0.610	**0.776**	0.615	0.440	0.294	0.702
LOC	P	0.748	0.508	0.745	0.554	0.445	0.241	0.676
	R	0.776	0.559	**0.777**	0.594	0.459	0.263	0.694
	F1	0.762	0.532	0.760	0.573	0.452	0.252	0.685
Micro Avg.	P	0.643	0.398	**0.650**	0.454	0.374	0.200	0.598
	R	0.633	0.383	0.627	0.411	0.356	0.171	0.580
	F1	0.638	0.390	**0.638**	0.431	0.365	0.184	0.589
MD	P	0.736	0.468	**0.744**	0.552	0.439	0.248	0.661
	R	0.725	0.451	0.717	0.500	0.417	0.211	0.642
	F1	0.730	0.459	**0.730**	0.525	0.428	0.228	0.652

Table 7: Breakdown of individual Chinese (ZH) results for all possible data augmentation methods (excluding group training as it was not grouped). Bold indicates if a metric performs better than or as good as the original.