Variability Across Languages in Zero-Shot Multilingual Learning

Manon Scholivet
Aix-Marseille Univ, Universite de Toulon, CNRS, LIS, Marseille, France
firstname.lastname@lis-lab.fr

Relevant UniDive working groups: WG3, WG4

1 Introduction

Annotations are data that can be expensive to obtain. However, many languages currently lack any annotation. In order to be able to process languages without annotation, we trained a tagger predicting the *Part Of Speech* (POS) of new languages never seen during training using annotations from other languages.

2 Experimental Settings

Tagger We used the tagger from (Dary and Nasr, 2021)[1], trained in three different configurations:

Mono corresponds to a training on a single language and tagging that same language for testing.

Multi represents a training on a set of 38 languages (see the list in Appendix A) and tagging each language using this single multilingual model.

ZS, for zero-shot, is identical to *Multi*, except that we exclude one language L. The evaluation of ZS is performed on the tagging of language L only. This experiment simulates a learning for a language for which we do not have training data.

Corpora The corpora are derived from the version 2.0 of *Universal Dependencies* (UD) (Nivre et al., 2016), balanced so that each language has 20,000 tokens for training. Test corpora have not been modified.

***World Atlas of Language Structures* (WALS)** For each language we work with, we extracted a vector of 22 features from WALS, which we will refer to as W_{22}. The list of features for this vector is in Appendix B.

3 Results

The complete results of all experiments are available in Appendix C. We observe that in a zero-shot ZS framework, the results vary significantly from one language to another, with a standard deviation of 17.06 between languages. This variation is much larger than the monolingual *Mono* experiments (2.72 standard deviation) or multilingual *Multi* experiments (3.23 standard deviation).

One hypothesis to explain this significant variability is the existence of a "close" language in the training set, which would allow for better knowledge sharing with the target language. Two ways to define a close language will be considered: a close language in the empirical sense and a close language defined based on a set of features from the WALS (Dryer and Haspelmath, 2013), by comparing the vectors of each language.

Existence of an empirically close language In a zero-shot setting, models are trained on the 38 languages minus one, the target language. Is the presence of a close language to the target language among the remaining 37 languages in the training set important? Could the variability in zero-shot conditions come from this?

The existence of a close language in the training set could help achieve better knowledge sharing with the target language. To calculate the existence of a close language empirically, we examine the results of a monolingual tagger *Mono* for language $L1$ applied to language $L2$.

We obtain a 38x41 matrix[2] to empirically estimate the proximity of languages to each other. We define a new measure, the Closest Language (CL) of language L, as the language whose model will give the best score when tagging language L. The result of the score for tagging language L by the CL gives us an empirical measure of the isolation of language L (see Table 1). The lower the result, the more isolated the language is.

To verify whether the CL score plays a role in the results of the ZS experiment, we measured the Pearson correlation between the scores of these two measures. We found a correlation of 0.95. The presence of a close language seems to be crucial

[1] https://gitlab.lis-lab.fr/franck.dary/macaon

[2] 38 languages with training data and thus a model, and 41 languages with test data (the three different languages are bxr, kmr, and sme).

Lang.	CL	CL Score	Lang.	CL	CL Score
ar	fa	53.00	it	es	66.46
bg	ru	71.63	ja	pl	39.99
bxr	cs	45.98	kmr	tr	38.12
ca	es	79.20	ko	cs	52.91
cs	sl	73.63	lv	sl	57.85
da	nob	80.12	nl	nob	54.61
de	nl	47.42	nno	nob	83.02
el	cs	44.74	nob	nno	84.07
en	fr	47.18	pl	cs	71.72
es	ca	82.20	pt	es	68.98
et	fi	62.67	ro	fr	55.91
eu	fi	52.73	ru	bg	76.50
fa	sl	54.94	sl	hr	66.54
fi	et	61.36	sme	fi	39.15
fr	ca	67.80	sv	da	73.60
ga	pl	38.87	tr	eu	53.71
he	cs	45.82	uk	ru	75.12
hi	sl	40.94	ur	sl	44.36
hr	sl	76.33	vi	ko	46.18
hu	pt	51.92	zh	ja	44.16
id	fi	56.12			

Table 1: CL for the 41 test languages. Languages in yellow belong to the Slavic language family, in red to the Germanic language family, and in blue to the Romance language family.

in explaining the results obtained in a zero-shot context, and the results of zero-shot experiments depend on the CL score.

Existence of a close language in terms of WALS Calculating the CL, although extremely useful, is also costly in terms of time and computation. Moreover, in extreme zero-shot conditions, where absolutely no annotated data exists, not even for evaluation, it is not possible to find the CL since decoding and evaluating a test corpus won't be possible. As the CL scores are highly correlated with zero-shot results, it provided an advance estimate of prediction accuracy. We would like to find, using the WALS, a measure of language isolation that is correlated with ZS results, similar to CL.

To quantify the isolation of a language, we define the *Connectedness Index* (CI) of a language as the average number of feature values it shares with other languages. The CI provides a measure of language isolation based on the WALS. This measure involves pairwise comparisons of the W_{22} vector of the language L with all other languages in a set, averaging the number of shared features with other languages:

$$CI(L) = \frac{100}{k} \sum_{f=1}^{k} \frac{1}{N-1} \sum_{L' \neq L} \delta(W(L', f), W(L, f))$$

where k is the dimension of the WALS vector,

Lang.	CI	Lang.	CI	Lang.	CI
ar	57.00	fr	58.85	nob	59.09
bg	63.64	ga	53.69	pl	64.25
ca	53.93	he	61.67	pt	66.34
cs	54.55	hi	48.03	ro	59.95
da	59.83	hr	64.37	ru	67.32
de	47.17	hu	53.19	sl	66.34
el	62.16	id	68.30	sv	59.83
en	65.11	it	48.03	tr	31.57
es	61.79	ja	30.71	uk	68.30
et	62.04	ko	41.65	ur	48.03
eu	39.07	lv	54.05	vi	57.49
fa	43.24	nl	47.17	zh	53.69
fi	57.13	nno	59.09		

Table 2: CI for W_{22} for the 38 training languages.

N[3] is the number of languages, $W(l, f)$ is the value of feature f for language L, and δ is the Kronecker delta[4]. $CI(L)$ indicates how much the WALS vector for language L shares its values with other languages. $CI(L) = 0$ is the case where the feature values of language L are not found in any other language. In the case where $CI(L) = 100$, L shares all feature values with all languages in the set. This situation occurs if all languages have identical vectors. The CI for W_{22} for the 38 training languages is available in Table 2.

When measuring the correlation between CI and ZS, a score of 0.50 is obtained. This score is significantly lower than the correlation of 0.95 obtained between the CL score and the ZS experiment scores. The WALS does not estimate the ZS score as effectively as the CL score. However, they are still a bit correlated, providing some indication of the predictions quality in zero-shot conditions.

4 Conclusions and Future Work

The variability in zero-shot predictions is mainly explained by the presence of a close language in the training corpus, which has a major impact on results for ZS-type experiments. The more isolated a language is, according to WALS but especially in the empirical sense, the lower the ZS results will be. The correlation with CI scores, although less significant, still provides an indication of the quality of zero-shot results and has the advantage of only requiring the WALS.

Additional experiments using WALS as input could help in knowledge sharing between languages. Using languages from the same language family could also be a lead for further exploration.

[3] here, $N = 38$, for the 38 languages in the training set

[4] $\delta(x, y) = 1$ if $x = y$ and 0 otherwise

References

Franck Dary and Alexis Nasr. 2021. The Reading Machine: a Versatile Framework for Studying Incremental Parsing Strategies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *LREC*.

A List of languages

Arabic (ar), Bulgarian (bg), Buryat (bxr), Catalan (ca), Czech (cs), Danish (da), German (de), Greek (el), English (en), Spanish (es), Estonian (et), Basque (eu), Farsi (fa), Finnish (fi), French (fr), Irish (ga), Hebrew (he), Hindi (hi), Croatian (hr), Hungarian (hu), Indonesian (id), Italian (it), Japanese (ja), Kurmanji (kmr), Korean (ko), Latvian (lv), Dutch (nl), Norwegian (no), Polish (pl), Portuguese (pt), Romanian (ro), Russian (ru), Slovenian (sl), Northern Sami (sme), Swedish (sv), Turkish (tr), Ukrainian (uk), Urdu (ur), Vietnamese (vi), Chinese (Mandarin) (zh).

The Buryat (bxr), Kurmaji (kmr) and Northern Sami (sme) did not have training sets. They were used only in the test set.

B Features from the WALS

- 81A : Order of Subject, Object and Verb
- 82A : Order of Subject and Verb
- 83A : Order of Object and Verb
- 85A : Order of Adposition and Noun Phrase
- 86A : Order of Genitive and Noun
- 87A : Order of Adjective and Noun
- 88A : Order of Demonstrative and Noun
- 89A : Order of Numeral and Noun
- 90A : Order of Relative Clause and Noun
- 92A : Position of Polar Question Particles
- 94A : Order of Adverbial Subordinator and Clause
- 95A : Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase
- 96A : Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun
- 97A : Relationship between the Order of Object and Verb and the Order of Adjective and Noun
- 101A : Expression of Pronominal Subjects
- 112A : Negative Morphemes
- 116A : Polar Questions
- 143A : Order of Negative Morpheme and Verb
- 143E : Preverbal Negative Morphemes
- 143F : Postverbal Negative Morphemes
- 143G : Minor morphological means of signaling negation
- 144A : Position of Negative Word With Respect to Subject, Object, and Verb

C Results

Lang.	*Mono*	*Multi*	*ZS*	Lang.	*Mono*	*Multi*	*ZS*
ar	92.94	92.03	60.23	bg	95.73	95.09	78.16
ca	95.84	96.06	83.48	cs	95.12	94.02	81.11
da	92.41	91.34	83.93	de	89.16	88.94	53.84
el	95.47	95.28	47.61	en	89.54	88.53	27.96
es	93.22	94.00	88.69	et	90.02	86.49	63.55
eu	90.62	88.16	55.73	fa	94.83	93.78	56.24
fi	84.90	82.25	67.08	fr	93.45	92.97	78.29
ga	89.11	87.02	45.28	he	94.45	93.84	52.30
hi	93.62	92.75	35.67	hr	94.30	93.38	83.07
hu	92.90	90.72	60.67	id	90.64	89.30	59.04
it	94.30	93.94	77.09	ja	92.41	91.78	33.28
ko	93.38	91.74	50.23	lv	89.44	85.84	62.75
nl	87.78	86.64	61.04	nno	90.43	90.00	84.77
nob	91.92	91.41	88.65	pl	95.19	93.46	78.82
pt	93.25	93.00	76.11	ro	94.11	92.23	64.28
ru	94.91	94.23	85.17	sl	92.50	89.48	69.10
sv	92.03	91.44	79.70	tr	90.44	86.70	56.34
uk	91.93	91.44	80.07	ur	90.56	89.53	38.97
vi	86.05	84.52	39.25	zh	87.98	88.16	50.51
Avg.	92.02	90.81	64.16				

Table 3: Accuracy of the POS prediction for each configuration and each language