Sharing Data by Language Family: Data Augmentation for Romance Language Morpheme Segmentation

Lauren Levine
Georgetown University
Department of Linguistics
lel76@georgetown.edu

Abstract

This paper presents a basic character level sequence-to-sequence approach to morpheme segmentation for the following Romance languages: French, Italian, and Spanish. We experiment with adding a small set of additional linguistic features, as well as with sharing training data between sister languages for morphological categories with low performance in single language base models. We find that while the additional linguistic features were generally not helpful in this instance, data augmentation between sister languages did help to raise the scores of some individual morphological categories, but did not consistently result in an overall improvement when considering the aggregate of the categories.

1 Introduction

Morpheme segmentation is a task in which individual words are divided into meaningful sub-units called morphemes. It is a difficult task, particularly in synthetic languages which have more complex morphological systems, but morphological analysis is an important sub-component of various downstream NLP related tasks, such as lexicography, terminology management, and semantic parsing. Previous approaches to morpheme segmentation include unsupervised methods (Creutz and Lagus, 2007), and more recently there have been neural approaches (Wang et al., 2016).

This paper is a submission to the SIGMORPHON 2022 shared task on morpheme segmentation, which aims to benefit the NLP community with improvements for subword-based tokenization through morpheme segmentation (Batsuren et al., 2022). The shared task includes word-level and sentence-level morpheme segmentation subtasks for various development languages. We focus on the subtask for word-level morpheme segmentation, specifically for the three Romance languages among the development languages: French, Italian, and Spanish. In this paper, we experiment with adding character based features to a sequence to sequence neural model, and we also experiment with sharing training data between sister languages.

The structure of the of the paper is as follows: In Section 2 we give an overview of the base system architecture of our approach. Section 3 describes the character based features we experimented with during development, and Section 4 describes our methods for data sharing between sister languages. Section 5 presents the results from our various models, and Section 6 provides the accompanying discussion. Finally, Section 7 offers a brief conclusion.

2 System Architecture[1]

We take a character-level sequence-to-sequence approach as the base architecture for our morpheme segmentation models. We base our approach on a simple recurrent model in the Keras[2] framework and adapted the base model to fit the needs of the word-level morpheme segmentation task. The encoder and decoder for the model each contain a single GRU layer. The batch size was 64 and the latent dimension of the encoding space was 256. All models were trained with early stopping with a max of 30 epochs. Base models for each of our focus languages (French, Italian, Spanish) were trained on this architecture using only the language specific training data provided by the shared task for the word-level subtask. The performance of these models is described in Section 5.2.

3 Additional Features

While sequence-to-sequence neural models have a tremendous ability to learn patterns that are la-

[1] https://github.com/lauren-lizzy-levine/2022SegmentationST.git

[2] https://keras.io/examples/nlp/lstm_seq2seq/

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 117 - 123
July 14, 2022 ©2022 Association for Computational Linguistics

tent in the raw text data on which the models are trained, there is still value in leveraging additional knowledge sources to provide features that may be linguistically important to morpheme segmentation that cannot be gleaned from the raw text of the training data alone. This is particularly true for languages where training data is limited and for morphological categories that are represented with low frequency in the data.

In order to train extra features in sequence to sequence modeling, we can combine our features into a single input vector with the individual input character representations (Sundaramoorthy, 2017). We do this by concatenating vectorized character input with a vectorized representation of our character based features. For simplicity's sake, we experimented with a series presence/absence features, which could be represented with a binary 1 or 0 encoding and easily concatenated to the one-hot representation of the text of the input character.

We experimented with adding a series of binary features to indicate whether the substrings that would be created by making a morpheme boundary at a given character would contain a known prefix or suffix. We created character based features for the following rules (Yes-1, No-0):

If the given character were the start of a new morpheme:

1. Is the string to the left of the boundary a prefix?

2. Is the string to the right of the boundary a suffix?

3. Does a substring to the left (ending at the morpheme boundary) contain a prefix?

4. Does a substring to the left (ending at the morpheme boundary) contain a suffix?

5. Does a substring to the right (starting from the morpheme boundary) contain a prefix?

6. Does a substring to the right (starting from the morpheme boundary) contain a suffix?

For instance, the word *enthrallments* would have the feature vector *000001* for the character *m*, as visualized in Figure 1. This is because *ment* is a known suffix that starts a character *m* where we are imagining a morpheme boundary to be, which means that "Yes" is the answer for question six. The answer for the rest of the questions is "No", so the rest of the digits in the vector are *0*.

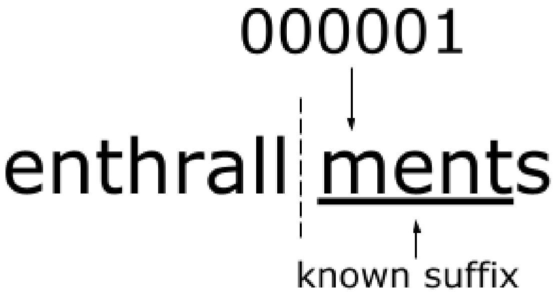

Figure 1: Visualization for the feature vector and corresponding potential morpheme boundary for the character *m* in the word *enthrallments*.

Such short feature vectors were generated for every character in every word of the provided data sets for our focus languages by referencing against previously complied language specific prefix[3] and suffix[4] lists compiled from Wiktionary.

We created various models with subsets of the training data and tested on subsets the development data for validation in order to gauge the merit of these features. In this instance, the inclusion of various combinations of the above features frequently led to degradation in performance compared to our base models when evaluated on the development data. As such, the features described above were not included in our final models trained on the full data set for most of our focus languages. For the sake of comparison, in Section 5.3 we include the results of a model trained on the full French training data which also incorporates a subset of the features outlined above. This model shows marginal improvement over the base French model on the test data.

4 Sister Language Data Sharing

Data augmentation for low-resource languages has been well researched area for various NLP tasks, such as machine translation (Fadaee et al., 2017) and speech recognition (Ragni et al., 2014). While data is provided by the shared task for all of the development languages, the number of training instances varies considerably, both in total amount and in the proportion of different morphological categories attested. Sharing data between languages is one means of evening out the representation of these underrepresented morphological cate-

[3] https://en.wiktionary.org/wiki/Category:Prefixes_by_language

[4] https://en.wiktionary.org/wiki/Category:Suffixes_by_language

Word Class	Description
`000`	Root words
`001`	Compound only
`010`	Derivation only
`011`	Derivation and Compound
`100`	Inflection only
`101`	Inflection and Compound
`110`	Inflection and Derivation
`111`	Inflection, Derivation, Compound

Table 1: Word class codes for morphological categories in training and development data.

gories. This type of data sharing is an instance of transductive transfer learning, where the domains are initially distinct (different languages), but the task in question remains the same (morpheme segmentation), and the knowledge in one domain is used to increase the task performance in the other domain (Pan and Yang, 2010).

Sister languages descend from a common ancestral language and are as such part of the same language family. Languages from the same language family are more likely to bear a strong resemblance to one another with regard to various linguistic aspects, including morphological structure, than sets of unrelated languages.

Our focus languages in this paper (French, Italian, and Spanish) are all a part of the Romance language family, and as such, we may posit that they share enough similarity in their morphological structure for there to be some benefit in sharing data between the languages during training.

In order to test this conjecture, we make a comparison between base models for each of our focus languages, which only contain training data from one language, and augmented models, which are trained on the full training data for one language and supplemented with training data from the other two Romance languages for select morphological categories.

For several of the development languages, including all three Romance languages, training and development data for the word-level subtask included additional annotation which indicated the morphological category of the word, and the evaluation scripts provided by that shared task also offered a breakdown by morphological category. The morphological categories provided in the shared task data are shown in Table 1.

In order to decide which morphological categories should be augmented with data from sister languages for each of the Romance languages, we evaluate our base models, which were each only trained with data from one language. For each language, we examine the base model's performance on the development data for the task, and we identify the four morphological categories with the lowest performance. For these categories, we add supplemental data from the other two Romance languages to train our augmented models. The identification of these categories for each of our augmented models and the results of their performance is detailed in Section 5.4.

5 Results

The shared task for word-level morpheme segmentation uses precision, recall, and F-measure as evaluation metrics for correctly predicted morphemes, as well as the average Levenshtein edit distance between the predicted instance and the reference instance. Overall scores are reported, as well as scores for individual morphological categories. The following subsections go through the baseline results provided by the shared task for our focus languages, as well as the results for our models. All scores are on the test data sets for individual languages. Overall, we find that all of our models make a significant improvement over the baseline.

5.1 Baseline

The baseline results given by the shared task for the Romance languages in the word-level subtask are all the results of Multilingual BERT Tokenizer (cased). Below are the overall baselines for French, Italian, and Spanish scored on the test data:

Lang.	P	R	F	Dist.
French	11.35	14.30	12.66	4.28
Italian	8.04	10.43	9.08	5.35
Spanish	15.59	17.68	16.57	5.21

5.2 Base Models

Base models for French, Italian, Spanish were trained on the architecture described in Section 2. Each model was trained on the entire training data for a single language. The results on the test data for each language broken down by morphological category are shown below. We note that these base models greatly outperform the baseline models from the previous sub-section.

French				
Cat.	P	R	F	Dist.
000	37.51	54.99	44.60	1.45
001	33.24	36.98	35.01	3.70
010	63.59	63.99	63.79	2.03
011	35.11	26.14	29.97	6.35
100	83.49	88.05	85.71	0.59
101	80.00	75.68	77.78	1.35
110	92.96	90.24	91.58	0.62
111	77.92	67.42	72.29	3.32
all	83.06	83.70	83.38	0.98

Italian				
Cat.	P	R	F	Dist.
000	39.94	57.44	47.12	1.53
001	23.40	22.92	23.16	4.27
010	71.93	71.99	71.96	1.68
011	32.43	26.67	29.27	6.43
100	84.04	88.18	86.06	0.64
101	47.56	42.86	45.09	4.80
110	93.86	91.28	92.55	0.60
111	48.39	31.91	38.46	6.27
all	87.21	87.77	87.49	0.78

Spanish				
Cat.	P	R	F	Dist.
000	44.16	61.50	51.41	1.23
001	13.11	13.79	13.45	4.72
010	68.93	65.43	67.13	1.59
011	36.36	21.05	26.67	7.67
100	95.27	96.25	95.76	0.23
101	86.24	77.25	81.50	1.31
110	98.35	97.32	97.83	0.18
111	93.67	86.05	89.70	2.00
all	96.00	95.90	95.95	0.27

Looking at the results above, we see that the relative performance on the different morphological categories amongst the three languages is relatively stable. All three of the languages have the highest scores on the *Inflection and Derivation (110)* category, followed by the *Inflection only (100)* category. For all three languages, the two lowest performing morphological categories are *Compound only (001)* and *Derivation and Compound (011)*.

We also note that the overall scores for each language relative to one another correlates to the size of the training data available: French has the least training data available, while Spanish has the most, and correspondingly, the overall scores for Spanish are the highest and the overall scores for French are the lowest. A table of the word category distributions within the shared task data for the three languages can be viewed in Appendix A. Predictions for all three of these models on the test data for their respective languages were submitted to the shared task (System GU-2).

5.3 Feature Model

As noted in Section 3, smaller trials during development indicated that the inclusion of the additional features we experimented with led to a degradation in performance. As such, we did not train a full set of feature models for all of our focus languages. For the sake of comparison, we trained a model on the full French training data with the first two features in our feature set:

If the given character were the start of a new morpheme:

1. Is the string to the left of the boundary a prefix?

2. Is the string to the right of the boundary a suffix?

The results for this model on the French development data are shown below. We note that in this instance there is marginal improvement when compared to the results of the French base model in the previous sub-section (gains/losses from the base model are listed in parentheses). Predictions from this model were not submitted to the shared task.

French	**with**	**Features**		
Cat.	P	R	F	Dist.
000	37.47 (-0.04)	56.09 (+1.10)	44.93 (+0.33)	1.51 (+0.06)
001	28.95 (-4.29)	32.54 (-4.44)	30.64 (-4.37)	3.90 (+0.40)
010	63.87 (+0.28)	64.95 (+0.96)	64.40 (+0.43)	2.00 (-0.03)
011	35.10 (-0.01)	30.11 (+4.97)	32.42 (+2.45)	6.35 (+0.00)
100	84.97 (+1.48)	89.12 (+1.07)	86.99 (+1.28)	0.56 (-0.03)
101	76.54 (-3.46)	68.24 (-7.44)	79.14 (+1.36)	1.83 (+0.48)
110	93.04 (+0.08)	90.16 (-0.08)	91.58 (+0.00)	0.60 (-0.02)
111	83.53 (+5.61)	79.78 (+12.36)	81.61 (+9.32)	2.21 (-1.11)
all	83.45 (+0.39)	84.13 (+0.43)	83.79 (+0.41)	0.96 (-0.02)

5.4 Augmented Models

The augmented models for each language were trained with additional data from the other two Romance languages. The morphological categories that were chosen to be augmented for each language were selected by identifying the lower performing morphological categories (bottom 4 categories) in the results of the base models on the development data for each language (listed in full in Appendix B). For selected categories, all of the training data from the other two Romance languages in those same categories was added to the training data of the original language to train the augmented model. For each language below, we identify the morphological categories that were augmented and list the results of the augmented model's performance on the test data of the original language (gains/losses from each language's respective base model are listed in parentheses). Predictions for the French and Italian models on the test data for their respective languages were submitted to the shared task (System GU-1).

French:

According to the results of the base model on the development data, the following categories had the lowest performance: *root words (000)*, *compound only (001)*, *derivation only (010)*, and *inflection only (011)*. The categories were augmented with Italian and Spanish training data from the same categories.

Cat.	P	R	F	Dist.
000	49.76 (+12.25)	67.40 (+12.41)	57.25 (+12.65)	1.03 (-0.42)
001	26.97 (-6.27)	28.40 (-8.58)	27.67 (-7.34)	3.99 (+0.29)
010	63.09 (-0.50)	61.71 (-2.28)	62.39 (-1.40)	1.98 (-0.05)
011	42.14 (+7.03)	33.52 (+7.38)	37.34 (+7.37)	5.45 (-0.90)
100	85.31 (+1.82)	88.90 (+0.85)	87.07 (+1.36)	0.53 (-0.06)
101	72.99 (-7.01)	67.57 (-8.11)	70.18 (-7.60)	1.83 (+0.48)
110	92.50 (-0.46)	89.39 (-0.85)	90.92 (-0.66)	0.62 (+0.00)
111	80.52 (+2.60)	69.66 (+2.24)	74.70 (+2.41)	3.00 (-0.32)
all	83.66 (+0.60)	83.21 (-0.49)	83.44 (+0.06)	0.93 (-0.05)

Comparing the above table to the base model results for French, we see that the augmented category *root words (000)* increases by the largest amount: +12.25 (P), +12.41 (R), +12.65 (F), -0.42 (Dist.). All of the scores for the other morphological categories either raise or fall by smaller margins. The sizable jump for *root words (000)* is likely do to the fact that it is a larger morphological class in the training data sets of our languages.

Italian:

According to the results of the base model on the development data, the following categories had the lowest performance: *compound only (001)*, *derivation and compound (011)*, *inflection and compound (101)*, and *inflection*, *derivation, compound (111)*. The categories were augmented with French and Spanish training data from the same categories.

Cat.	P	R	F	Dist.
000	42.75 (+2.81)	60.93 (+3.49)	50.25 (+3.13)	1.42 (-0.11)
001	18.00 (-5.40)	18.75 (-4.17)	18.37 (-4.79)	4.48 (+0.21)
010	73.48 (+1.55)	74.32 (+2.33)	73.90 (+1.94)	1.56 (-0.12)
011	34.21 (+1.78)	28.89 (+2.22)	31.33 (+2.06)	6.07 (-0.36)
100	85.67 (+1.63)	89.48 (+1.30)	87.54 (+1.48)	0.57 (-0.07)
101	54.02 (+6.46)	51.65 (+8.79)	52.81 (+7.72)	3.37 (-1.43)
110	94.68 (+0.82)	92.14 (+0.86)	93.39 (+0.84)	0.53 (-0.07)
111	63.64 (+15.25)	44.68 (+12.77)	52.50 (+14.04)	5.64 (-0.63)
all	88.41 (+1.20)	88.97 (+1.20)	88.69 (+1.20)	0.70 (-0.08)

Comparing the above table to the base model results for Italian, we see that the overall results increase by a small margin: +1.20 (P), +1.20 (R), +1.20 (F), -0.08 (Dist.). All of the morphological categories had slight increases from the base model, except for the *compound only (001)* category.

Spanish:

According to the results of the base model on the development data, the following categories had the lowest performance: *root words (000)*, *compound only (001)*, *derivation only (010)*, and *inflection only (011)*. The categories were augmented with French and Italian training data from the same categories.

Cat.	P	R	F	Dist.
000	59.24 (+15.08)	75.03 (+13.53)	66.21 (+14.80)	0.82 (-0.41)
001	9.09 (-4.02)	8.62 (-5.17)	8.85 (-4.60)	3.97 (-0.75)
010	65.12 (-3.81)	59.96 (-5.47)	62.43 (-4.70)	1.72 (+0.13)
011	35.71 (-0.65)	26.32 (+5.27)	30.30 (+3.63)	7.33 (-0.34)
100	94.79 (-0.48)	95.69 (-0.56)	95.24 (-0.52)	0.24 (+0.01)
101	81.38 (-4.87)	72.51 (-4.74)	76.69 (-4.81)	1.68 (+0.37)
110	98.06 (-0.29)	96.86 (-0.46)	97.46 (-0.37)	0.20 (+0.02)
111	92.31 (-1.36)	83.72 (-2.33)	87.80 (-1.90)	2.22 (+0.22)
all	95.72 (-0.28)	95.35 (-0.55)	95.53 (-0.42)	0.29 (+0.02)

Comparing the above table to the base model results for Spanish, we see that the augmented category *root words (000)* increases by a notable amount: +15.08 (P), +13.53 (R), +14.80 (F), -0.41 (Dist.). All of the scores for the other morphological categories fall by a notable margin. The sizable jump for *root words (000)* is likely do to the fact that it is a larger morphological class in the training data sets of our languages. The gains from the *root words* category do not balance out the losses from the other morphological classes, and we see a loss in the overall scores.

6 Discussion

While all of base models made significant improvements from the baseline scores provided for the word-level subtask, we note that our additional experimentation resulted in only modest improvements. We also note that our experimenting with additional features frequently led to score degradation on the development data.

We did not expect to see the general degradation in our scores with the inclusion of the known affix presence/absence based features that we saw in our experiments predicting on the development data. However, we did see the marginal improvement we expected in the results of the fully trained French model predicting on the test data, as described in Section 5.3. On possible explanation for these inconsistent results is that the inclusion of single character or two character affixes created feature vectors with too many false positives to be of use in the model's learning for our small scales experiments predicting on the development data. Further error analysis is needed to conclude the reason for such inconsistency. The fact that the improvements seen in fully trained French model were marginal suggest that the base architecture of our models may be independently capable of learning information encoded in our linguistic features.

The sharing of language data between sister languages gave modest gains in our experiments, indicating that there is some potential to leverage available data from morphologically similar languages for morpheme segmentation. In future experiments we want to experiment with different methods of deciding what/how much data should be shared in order to maximize this potential. Additionally, rather than just assuming that being in the same language family indicates enough morphological similarity between languages for data sharing to be of use, we believe that is would be beneficial to make a closer study of the morphological similarities and differences between sets of languages that will be used for data sharing.

7 Conclusion

In this paper we presented a basic approach to morpheme segmentation at the word-level for the SIGMORPHON 2022 shared task for French, Italian, and Spanish. All of our presented models considerably improved upon the baselines for the shared task. While the extra character based features we experimented with generally did not prove useful in this instance, we did find some evidence that sharing data between morphologically similar languages could result in minor improvements in the segmentation of words in morphological categories which were augmented with additional data.

References

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinović, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The sigmorphon 2022 shared task on morpheme segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Anton Ragni, Kate M Knill, Shakti P Rath, and Mark JF Gales. 2014. Data augmentation for low resource languages. In *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association*, pages 810–814. International Speech Communication Association (ISCA).

Shiva Sundaramoorthy. 2017. A novel approach to feed and train extra features in seq2seq (tensorflow amp; pytorch).

Linlin Wang, Zhu Cao, Yu Xia, and Gerard De Melo. 2016. Morphological segmentation with window lstm neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

A Language Data Statistics (word counts)

Word Class	French	Italian	Spanish
000	13619	21037	15843
001	1684	431	248
010	67983	41092	18449
011	506	140	82
100	105192	253455	502229
101	478	317	458
110	126196	237104	346862
111	186	158	343
Total Words	382797	553734	884514

B Performance of Base Models on the Development Data

French				
Cat.	P	R	F	Dist.
000	36.56	54.63	43.80	1.45
001	32.61	36.01	34.23	3.46
010	63.28	63.48	63.38	2.07
011	29.58	24.56	26.84	6.67
100	84.21	88.54	86.32	0.57
101	85.14	82.89	84.00	0.79
110	92.99	90.25	91.60	0.61
111	83.13	78.41	80.70	2.11
all	83.18	83.75	83.47	0.97

Italian				
Cat.	P	R	F	Dist.
000	43.08	60.93	50.47	1.40
001	25.26	25.53	25.40	4.06
010	70.97	71.73	71.35	1.73
011	26.67	27.91	27.27	6.07
100	84.15	88.18	86.12	0.64
101	56.96	49.45	52.94	3.07
110	93.84	91.16	92.48	0.60
111	66.67	55.32	60.47	4.64
all	87.21	87.75	87.48	0.78

Spanish				
Cat.	P	R	F	Dist.
000	43.87	60.82	50.97	1.24
001	15.79	15.52	15.65	3.34
010	67.63	64.62	66.09	1.67
011	18.18	10.53	13.33	5.17
100	95.32	96.27	95.79	0.23
101	87.23	80.79	83.89	1.10
110	98.36	97.30	97.83	0.18
111	86.67	77.38	81.76	2.44
all	95.99	95.87	95.93	0.27