Czech Grammar Agreement Dataset for Evaluation of Language Models

Vít Baisa

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xbaisa@fi.muni.cz

Abstract. AGREE is a dataset and task for evaluation of language models based on grammar agreement in Czech. The dataset consists of sentences with marked suffixes of past tense verbs. The task is to choose the right verb suffix which depends on gender, number and animacy of subject. It is challenging for language models because 1) Czech is morphologically rich, 2) it has relatively free word order, 3) high out-of-vocabulary (OOV) ratio, 4) predicate and subject can be far from each other, 5) subjects can be unexpressed and 6) various semantic rules may apply. The task provides a straightforward and easily reproducible way of evaluating language models on a morphologically rich language.

Key words: language model, grammar agreement, verb suffix, Czech, subject, predicate, dataset, evaluation, perplexity

1 Introduction

Language modeling is one of the most important research fields within natural language processing. Language models are used in many applications ranging from basic tasks as spell checking, diacritic restoration to complex tasks as automatic speech recognition and machine translation.

The vast majority of the improvements have been demonstrated on English since the mainstream datasets are: 1) Brown corpus [1], 2) Penn Treebank [2], 3) Wall Street Journal, 4) English Gigaword [3] and 5) One billion word benchmark (OBB) [4]. All of based on English data.

It can be shown that English is especially suitable for various techniques as it has quite poor morphology and low OOV rate resulting in the fact that relatively small vocabularies are sufficient to cover the vast majority of unseen data.

AGREE task is intended to provide an extrinsic way of evaluating language models on non-English language data. Despite the fact that the Czech subject-predicate agreement is exhibited by verb suffixes, the task is suitable to evaluate both word-based and character-based language models.

The task was inspired by Microsoft Research Sentence Completion Challenge (MSCC) task [5] which contains 1,040 sentences from five Sherlock

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 63–67, 2016.

Holmes novels by Sir A. C. Doyle and in each sentence, one word is missing and 5 alternatives are supplied. The task is deliberately hard for n-gram models as the missing words are sometimes impossible to guess from a local context.

2 AGREE task

In past tense verbs, subject-predicate agreement is exhibited by grammar suffixes *a*, *o*, *i*, *y* and an empty suffix. These correspond to the following subject types: 1) **-a**, e.g. *žila*: subject is feminine singular (she lived) or neuter plural (kittens lived), 2) **-o**, e.g. *žilo*: neuter singular (a pig lived), 3) **-i**, e.g. *žili*: masculine plural (men lived) or subject is a group of entities of feminine and masculine genders (men and women lived), 4) **-y**, e.g. *žily*: feminine plural (women lived), neuter plural (children lived), 5) **-∅**, žil: masculine singular (he lived).

Other rules mapy apply, e.g. a masculine animate subject outweighs inanimate subjects *muži a stroje pracovali* (men and machines worked) etc.

The nature of the task is similar to MSCC: to choose a suffix properly, the whole sentence must be comprehended. Even though the agreement is grammar-motivated, it depends on the semantics. Sometimes, even the whole sentence might be insufficient to choose a proper suffix.

Unlike in MSCC task, sentences might contain more than one position where a suffix (word) is to be chosen. Commented example sentences with marked verbs follow.

> *Pestrý program byl*** vítanou inspirací pro naše soubory.*

In this sentence, the subject *program* (programme) is governed by the predicate *byl* (was). To choose the right word form, language models need to capture just two neighbouring words—bigram.

> *Na zpáteční cestě do USA měl*** konvoj vézt zlato, platbu za dodané zbrojní zakázky.*

In this sentence it is enough to look at the subsequent word *konvoj* (convoy). Since Czech has rich morphology, it has free word order so the relative positions of predicates and subjects may vary.

> *Určitě tady všichni nešťastnému dědulovi drželi*** palce, ale to bylo*** asi všechno, co pro něho mohli*** udělat.*

Here the first verb is governed by word *všichni* (all of us), which is not neighbouring the verb in past tense. The second is related to the previous word (pronoun *to* (it) is a sign of an anaphora, the anaphoric subject here is the act of keeping one's fingers crossed). The gender of the pronoun is neuter and it is trivial to choose the right word form as the subject and predicate are next to each other. The third is again governed by the main subject (*všichni*). So the agreement is defined by a dependency spanning 12 words.

*Léon Bourgeois navrhl*** i praktický program solidarity.*

In this sentence, the gender of *Léon Bourgeois* needs to be guessed to assign a proper suffix. In the case of masculine, the suffix is empty and in the case of feminine it would be *-a*, i.e. *navrhla*. By omitting low frequency words (*Bourgeois*) which is a standard strategy in word-based techniques, models would probably miss this token and would be unable to learn what verbs with what endings co-occur with it.

*Teď už se normálně postavil*** a ťapkal*** trávou směrem ke mně.*

Some sentences are hard to complete without a proper context. The good strategy here is to assign the same suffix to all verbs since they are usually governed by the same (though sometimes unexpressed) subject. Nevertheless, the sentence indicates that the subject will be something like a cub, a puppy or a kitten. In Czech, puppies and cubs have usually neuter gender. But it might also be a named pet and in this case its real sex is determining the suffix.

*Ve třetím kole narazily*** na celkově třetí Třešňákovou s Pilátovou a podlehly*** jim 0 : 2 (-18, -8).*

Sometimes, common sense might help to guess the right word form. Here it is more probable that the gender of the unexpressed subject is feminine, as the sentence talks about a player couple facing some female opponents (we know it because of the surname endings *-ovou*) in a sport match. It is probable that the players will have the same gender as their opponents as it is usual in sports.

3 AGREE dataset

The dataset consists of 10 million Czech sentences from a Czech Web Corpus [6] with marked verbs in past tense split into three parts: 1) TRAIN with 9,900,000 sentences, 2) VALID with 99,000 sentences and 3) TEST with 996 sentences.

4 Evaluation of various models and a baseline

For the evaluation purpose, all possible combinations are generated yielding 17,940 sentences. TEST set has been manually annotated by several undergraduate students to estimate the difficulty of the task. The average verb accuracy was 86.5.

The baseline model chooses the most frequent word form for each marked verb (the frequencies were extracted from a Czech web corpus). Table 1 contains the summary of accuracies of baseline, random and various word- and character-based models.

The best result (59.8%) has been achieved by a word-based RNN model.[1] RNN performs only slightly better than SRILM 4-gram word-based model [7];

[1] `https://github.com/yandex/faster-rnnlm`

Table 1: The summary of AGREE task results for various models.

Model	Accuracy
Human (average)	86.5
Recurrent Neural Network with hidden layer size 100	59.8
SRILM word-based 4-gram	59.6
Chunk-based language model	58.7
SRILM character-based 9-gram	53.9
Baseline (the most frequent wordform)	42.0
Random (average of 10 runs)	19.6

both with standard settings. The vocabulary was not trimmed: the OOV of testing data against training data caused that in 1,401 sentences (out of 17,940) there was at least one OOV word. This might cause the rather poor results for word-based models .

The character n-gram model has been trained with SRILM, the chunk-based language model operating on byte level is described in [8].

5 Conclusion

AGREE task was developed to promote evaluation of language models focusing on morphologically rich languages. The task is motivated by a morphological phenomenon in Czech language of subject-predicate grammar agreement.

The task is hard as sentences must be comprehended and sometimes even common sense is needed for assigning the suffixes correctly.

Random choice yields 20%, baseline around 40%, best language models achieve 60% and human annotators 90%.

The dataset and auxiliary scripts have been released under Creative Commons Share-alike Attribution licence.[2]

In future we would like to build a similar task for even more morphologically rich languages where the OOV is more pronounced, e.g. Estonian, Hungarian and Turkish.

Acknowledgments. This work has been partly supported by the Masaryk University within the project *Čeština v jednotě synchronie a diachronie – 2016* (MUNI/A/0863/2015). The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

References

1. Francis, W.N., Kucera, H.: Brown corpus manual. Brown University (1979)

[2] `https://nlp.fi.muni.cz/~xbaisa/agree/`

2. Marcus, M., Marcinkiewicz, M., Santorini, B.: Building a large annotated corpus of English: The Penn Treebank. Computational linguistics **19**(2) (1993) 313–330
3. Graff, D., Kong, J., Chen, K., Maeda, K.: English Gigaword. Linguistic Data Consortium, Philadelphia (2003)
4. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., Robinson, T.: One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005 (2013)
5. Zweig, G., Burges, C.J.: The Microsoft Research sentence completion challenge. Technical report, Technical Report MSR-TR-2011-129, Microsoft (2011)
6. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The tenten corpus family. In: 7th International Corpus Linguistics Conference. (2013) 125–127
7. Stolcke, A.: SRILM-an extensible language modeling toolkit. In: Proceedings of the international conference on spoken language processing. Volume 2., Citeseer (2002) 901–904
8. Baisa, V.: Byte level language models. PhD thesis, Masaryk University (11 2016)