Classification of Language Proficiency Levels in Swedish Learners' Texts

Ildikó Pilán, Elena Volodina

Språkbanken, Department of Swedish, University of Gothenburg
Gothenburg, Sweden
ildiko.pilan@svenska.gu.se, elena.volodina@svenska.gu.se

Abstract

We evaluate a system for the automatic classification of texts written by learners of Swedish as a second language into levels of language proficiency. Since the amount of available annotated learner essay data for our target language is rather small, we explore also the potentials of domain adaptation for this task. The additional domain consists of coursebook texts written by experts for learners. We find that already with a smaller amount of in-domain Swedish learner essay data it is possible to obtain results that compare well to state-of-the-art systems for other languages, with domain adaptation methods yielding a slight improvement.

1. Introduction

In a second and foreign language (L2) teaching scenario, a longer piece of learner-written text is a wide-spread means to assess learners' level of progression (*proficiency level*). The human assessment of such texts, however, requires considerable time and effort. Moreover, such an assessment is prone to subjectivity: a negative attitude to a learner, hunger or bad mood may influence the decision. Therefore, there has been an increasing trend to complement human assessment with more objective and efficient computerized systems, e.g. Burstein (2003).

The automatic assessment of learner texts for Swedish has been investigated in the context of native language (L1) users in Östling et al. (2013). However, the classification of learner texts into L2 proficiency levels has not been previously explored for the same target language. For a successful L2 instruction and assessment, taking into consideration proficiency levels is crucial since beginners and more advanced learners show significant differences in terms of the lexical and grammatical patterns that they are able to process and produce.

We propose a machine learning based system for the automatic classification of Swedish L2 learner essays into proficiency levels. Since the collection and annotation of such texts presents a number of challenges (e.g. obtaining permits, anonymization, digitization), the amount of available annotated data can be rather limited. Therefore, besides proposing a model trained only on the small amount of available essay data for L2 Swedish, we investigate also the usefulness of domain adaptation methods for this task. As additional domain we use expert-written reading passages intended for learners as practice material. Transferring models between these domains has not been previously tested, however, since such texts can also be divided into the same type of learning levels, we hypothesize that they may constitute a useful source of additional training instances for classifying L2 essays. Our experiments show that training on the small L2 essay data achieves an accuracy of 72.2%, a performance similar to the state-of-the-art results reported for other languages for this task. Furthermore, some domain adaptation methods yield some improvement over this score.

2. Background

Learner-written texts can be assessed either in terms of a grade within a pass-fail range or in terms of levels of progression, e.g. school grade levels or L2 proficiency levels. A popular scale for L2 levels is the CEFR, i.e. the Common European Framework of Reference for Languages (Council of Europe, 2001). The CEFR defines language proficiency across 6 levels: A1 (beginner), A2, B1, B2, C1, C2 (proficient user), and it is wide-spread not only in Europe but also outside.

In recent years, a number of studies have appeared about CEFR level classification, but primarily for expert-written texts. For this text type such a classification has also been referred to as *L2 readability* and it has been explored for e.g. English (Xia et al., 2016), French (François and Fairon, 2012) and Chinese (Sung et al., 2015). Classifying CEFR levels in learner texts has been investigated to a lesser extent. Previous work of this kind includes Hancke and Meurers (2013) for German and Vajjala and Lõo (2014) for Estonian. An essay grading system with the use of domain adaptation is presented in Zesch et al. (2015). The authors conclude that a model can be successfully transfered between two different writing tasks when certain domain-specific features are excluded.

For Swedish, in terms of the type of data used, a related previous work is the L1 essay grading system presented in Östling et al. (2013). Moreover, in Pilán et al. (2015) we described a system classifying CEFR levels using L2 Swedish coursebook texts.

3. Datasets

Our datasets consist, on the one hand, of essays written by learners of L2 Swedish from the SweLL corpus (Volodina et al., 2016) and, on the other, by expert-written texts from the COCTAILL corpus (Volodina et al., 2014) intended for learners. From the latter source only reading comprehension texts were included. Both text types were manually annotated for CEFR levels and automatically annotated for linguistic elements such as parts of speech and dependency relations using the Sparv pipeline (Borin et al., 2012). Table 1 shows the amount of data per type and level.

Writer	Unit	A2	B1	B2	C1	Total
Learner	**Texts**	83	75	74	88	**320**
	Tokens	18K	29K	32K	60K	**140K**
Expert	**Texts**	157	258	288	115	**818**
	Tokens	37K	79K	101K	71K	**289K**

Table 1: CEFR-level annotated Swedish datasets.

4. Features

We used the feature set presented in Pilán et al. (2015) created for the assessment of readability in expert-written texts for L2 Swedish learners. The features span five dimensions: length-based (e.g. sentence and token length), lexical (e.g. word frequencies, CEFR level per token), morphological (e.g. ratio of past tense verbs to all verbs), syntactic (e.g. average length of dependency arcs) and semantic features (e.g. average number of word senses). For a detailed description of the features see Pilán et al. (2015).

5. Method

Instead of using data with the same distribution of feature values for both training and testing, in a domain adaptation setup we transfer knowledge from a *source domain* (D_S), the L2 coursebook texts in our case, to a *target domain* (D_T), the learner-written texts.

Our two baselines consist of the most frequent label in the dataset (MAJORITY) and an IN-DOMAIN setup with the learner essays as training as well as test data in a cross-validation setting. We compare these to different domain adaptation setups inspired by Daumé III and Marcu (2006).

In the SOURCE-ONLY setup, a model trained on all available coursebook texts was used to predict levels in the learner essays. In the COMBINED and WEIGHTED setups the training data includes also 60% of the D_T besides the D_S instances, in the latter case D_T instances receiving a higher weight. These setups vary in terms of the amount of informing instances, i.e. texts from which information is incorporated in the models. The least data intensive scenario is IN-DOMAIN based only on 320 learner essays, followed by SOURCE-ONLY using 818 coursebook texts without the prerequisite of annotated learner-written texts.

The classification models in all cases were based on the sequential minimal optimization (SMO) algorithm from WEKA (Hall et al., 2009) with the default parameter settings, and the feature set mentioned in Section 4. We report results in terms of F_1 score and quadratic weighted kappa (κ^2), a distance-based scoring function that takes into account also the degree of misclassifications.

6. Results and Discussion

The results of our experiments with and without domain adaptation are presented in Table 2. The IN-DOMAIN baseline using the limited amount of learner-written texts with 10-fold cross-validation was .721 F_1 and .886 κ^2. Employing a model based only on the coursebook texts (SOURCE-ONLY) for classifying the essays resulted in a considerably lower performance (-.283 F_1 and -.173 κ^2). Adding information from the 818 instances consisting of expert-written texts in COMBINED and WEIGHTED improved somewhat over the IN-DOMAIN setup.

	F_1	κ^2
MAJORITY	.120	.000
IN-DOMAIN	.721	.886
SOURCE-ONLY	.438	.713
COMBINED	.733	.863
WEIGHTED	**.747**	**.890**

Table 2: In- and cross-domain experiment results.

Predictions				
A2	**B1**	**B2**	**C1**	
27	1	0	0	**A2 instances**
2	22	7	1	**B1 instances**
0	5	17	6	**B2 instances**
0	0	11	29	**C1 instances**

Table 3: Confusion matrix for WEIGHTED.

The confusion matrix for WEIGHTED for the held-out 40% of the essay data is presented in Table 3. With WEIGHTED, all errors except one lied within a distance of one CEFR level compared to the annotated labels. In all setups lexical features, i.e. the amount of tokens per CEFR level and word frequency information were among the strongest predictors.

Our system performs well compared to in-domain systems for other languages using supervised machine learning methods. Our in-domain model achieved an accuracy of 72.2%, while Hancke and Meurers (2013) report 64.5% accuracy for a 5-level CEFR classification of German L2 learner essays. Vajjala and Lõo (2014), on the other hand, obtain 79% accuracy on Estonian L2 learner texts when classifying four CEFR levels (A2-C1). It is worth noting that in these two experiments about three times more in-domain training data was available.

7. Conclusion

We described a system for classifying proficiency (CEFR) levels in texts written by L2 Swedish learners. We experimented with both in-and cross-domain machine learning methods and found that they achieve a similar performance, both of which compare well to the state of the art for this task. In the future, additional domain adaptation methods could be explored to improve the trade-off between performance gain and the annotation cost of the additional out-of-domain data. Moreover, since learner texts are error-prone whilst expert-written texts are relatively error-free, the effects of automatic error correction could be investigated on the successfulness of the domain transfer.

References

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp - the corpus infrastructure of Språkbanken. In *LREC*, pages 474–478.

Jill Burstein. 2003. The e-rater Scoring Engine: Automated Essay Scoring with Natural Language Processing. *Lawrence Erlbaum Associates, Inc.*

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126.

Thomas François and Cédrick Fairon. 2012. An 'AI readability' formula for French as a foreign language. In *Proceedings of the EMNLP and CoNLL 2012*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research (LCR) conference*.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *To appear in International Journal of Computational Linguistics and Applications*. Available at `http://arxiv.org/abs/1603.08868`.

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling l2 texts through readability: Combining multilevel linguistic features with the cefr. *The Modern Language Journal*, 99(2):371–391.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR Level Prediction for Estonian Learner Text. *NEALT Proceedings Series Vol. 22*, pages 113–127.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *NEALT Proceedings Series Vol. 22*, pages 128–144.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. Swell on the rise: Swedish learner language corpus for european reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA, June. Association for Computational Linguistics.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-Independent Features for Automated Essay Grading. In *Proceedings of the Building Educational Applications Workshop at NAACL*.

Robert Östling, André Smolentzov, Björn Tyrefors, and Erik Höglin. 2013. Automated Essay Scoring for Swedish. In *The 8th Workshop on Innovative Use of NLP for Building Educational Applications*.