DDisCo: A Discourse Coherence Dataset for Danish

Linea Flansmose[1*], Oliver Kinch[2], Anders Jess Pedersen[2], Ophélie Lacroix[3*]

[1]Aarhus University, Langelandsgade 139, 8000 Aarhus, linea.flansmose@gmail.com; [2]Alexandra Institute, Rued Langgaards Vej 7, 2300 Copenhagen, {oliver.kinch, anders.j.pedersen}@alexandra.dk; [3]Wunderman Thompson MAP, Glentevej 61, 2400 Copenhagen, ophelie.lacroix@wundermanthompson.com;
[*]Research conducted at the Alexandra Institute.

Objective

Present DDisCo, a dataset including text from the Danish Wikipedia and Reddit annotated for discourse coherence. DDisCois an annotated dataset consisting of real-world texts, contrary to artificially incoherent text for training and testing models.
Presentation of performance and evaluation of several methods, including neural networks, on the dataset.

Data - Collection and Annotation

The data collected for this project includes: blog posts from the Reddit forum and encyclopedic texts from the Danish Wikipedia. This data was chosen with some ideals in mind: the texts should be written by a *variety of people*; the texts should not be *edited by professionals*; the texts should be of a *certain length*; the dataset should ideally show texts of *low, medium and high coherence*; the data could be made *publicly available* under a licence that allows commercial use.
The texts were annotated for coherence on a 3-points Likert scale: *low coherence*, *medium coherence*, *high coherence*. Following guidelines from [1, 4, 9], texts are considered *lowly coherent* when they are difficult to understand, unorganized, contained unnecessary details and can not be summarized briefly and easily, and vice versa for highly coherent.

Domain	Train	Test	Total
Reddit	401	100	501
Wikipedia	400	100	500
All	801	200	1001

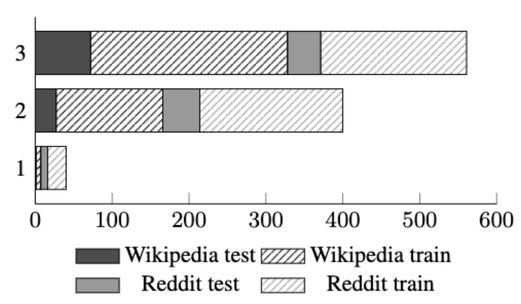

Table 1. Number of texts in the DDisCo dataset.
Figure 1. Distribution of coherence ratings in the dataset.

Feature-, and Text-based Classification

Feature-based Classification

The feature-based strategy consists in pre-calculating relevant numerical features and using these as input. We choose to compare the following four algorithms: Multinomial Näive Bayes (**NB**), Support Vector Machine (**SVM**), Random Forest (**RF**) and Logistic Regression (**LR**). The numerical features are the following:

- **LIX** [2], a readability index adapted to Danish.
- A weighted score derived from **the entity graph** [7], which is a measure of local coherence in a text.
- The number of **conjunctions** for each text. Conjunctions are markers of cohesion which are predominant indicators of coherence [8].

Text-based Classification

In the text-based strategy, the text is directly transformed into an embedding using different preprocessing methods and then fed to a machine or deep learning algorithm for training. We consider NB, SVM, RF, LR for the following embeddings:

- TF-IDF vectorizer with unigrams, bigrams and trigrams.
- (Facebook) Danish word embeddings [3].

We fine-tune several transformer-based pre-trained models for discourse coherence classification:

- daBERT (i.e. Nordic BERT): a BERT-based [6] model pre-trained on danish texts;
- mBERT: a multilingual BERT-based pretrained model;
- XLM-R: a multilingual XLM-Roberta-based [5] pre-trained model.

Experiments

The baseline (*Majority*) strategy represents a model that would always predict the most common rating. Each other score is an average on 5 runs. For each experiment, we split the training dataset randomly (80% train, 20% develop). For the feature-based strategy, we report only the results of the best classifier. For the text-based strategy with machine learning algorithms, we report the result of each classifier but only the one with the best text pre-processing strategy (lemmas or word embeddings).

Results

Table 2 shows the performance of the different models. Globally, the deep learning models achieve the best scores. Among the feature-based models, the conjunction feature is the most relevant for predicting discourse coherence ratings.

Input	Model	Acc.	Prec.	Rec.	F_1
Baseline					
-	Majority	0.57	0.32	0.57	0.41
Feature-based					
LIX	RF	0.49	0.50	0.49	0.49
EGraph	RF	0.50	0.50	0.50	0.50
Conj.	RF	0.59	*0.55*	0.59	0.53
All feats	NB	*0.60*	*0.55*	*0.60*	*0.56*
Text-based ML					
Lemmas	LR	0.58	0.33	0.58	0.42
Lemmas	SVM	0.63	0.59	0.63	*0.58*
Lemmas	NB	*0.64*	*0.61*	*0.64*	*0.58*
WV	RF	0.60	0.56	0.60	0.57
Text-based DL (transformers)					
Text	daBERT	0.65	0.61	0.65	0.62
Text	mBERT	**0.67**	**0.64**	**0.67**	**0.63**
Text	XLM-R	0.66	0.63	0.66	**0.63**

Table 2. Discourse coherence results, i.e. accuracy (Acc.), recall (Rec.), precision (Pre.) and weighted F_1 score. Inputs: Word vectors (WV). Scores in italic are the highest within the same strategy. Scores in bold are the highest globally.

References

[1] R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008. doi: 10.1162/coli.2008.34.1.1. URL `https://aclanthology.org/J08-1001`.

[2] C. Björnsson. Læsbarhed, 1968.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL `https://aclanthology.org/Q17-1010`.

[4] J. Burstein, J. Tetreault, and M. Chodorow. Holistic discourse coherence annotation for noisy essay writing. *Dialogue & Discourse*, 4(2):34–52, 2013.

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[7] C. Guinaudeau and M. Strube. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL `https://aclanthology.org/P13-1010`.

[8] M. A. K. Halliday and R. Hasan. *Cohesion in english*. Number 9. Routledge, 2014.

[9] A. Lai and J. Tetreault. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5023. URL `https://aclanthology.org/W18-5023`.