INSTITUTIONEN FÖR
SVENSKA SPRÅKET

GU-ISS-2020-04

SwedishGLUE – Towards a Swedish Test Set for Evaluating Natural Language Understanding Models

Yvonne Adesam, Aleksandrs Berdicevskis och Felix Morger

Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet
Research Reports from the Department of Swedish

ISSN 1401-5919

SwedishGLUE – Towards a Swedish Test Set for Evaluating Natural Language Understanding Models

Yvonne Adesam Aleksandrs Berdicevskis Felix Morger

Språkbanken, Department of Swedish,
University of Gothenburg

December 11, 2020

1 Introduction

GLUE[1] and SuperGLUE[2] are collections of tasks for evaluating Natural Language Understanding (NLU) models (Wang et al., 2018, 2019). They consist of a number of pre-existing (possibly adapted) datasets, and can be used for benchmarking systems, with the explicit goal of rewarding general systems that can handle different linguistic tasks across different domains. GLUE/SuperGLUE currently only deal with English data, and the goal of the SwedishGLUE project (funded by Vinnova 2020–2021, ref: 2020-02523) is to create an evaluation set for Swedish.

The GLUE/SuperGLUE data consist of datasets that were already available. Some of them were already used for training and evaluating systems on a particular task, some have been adapted to be more suitable for evaluation. Generally, while Swedish is in no way a low-resource language, the amount of language resources available for English is unsurpassable. The first steps for developing a Swedish evaluation set are therefore to determine which tasks should be prioritized for evaluation, and making an inventory of which appropriate data are available for Swedish, and which could easily be created, or converted into a fitting form. Larger sets of training data are not within the scope of the current project.

The SwedishGLUE project application states that the following should be part of the Swedish evaluation data: sentence-level semantic similarity, words in context, bias, reading comprehension, and inference. Highest priority are reading comprehension (cf. Multi-Sentence Reading Comprehension and BoolQ

[1] See `https://gluebenchmark.com`

[2] See `https://super.gluebenchmark.com`

from SuperGLUE), bias (cf. Winogender schemas/AX-g from SuperGLUE), and semantic similarity (cf. Microsoft Research Paraphrase Corpus from GLUE).

2 Desiderata

In this section, we describe what requirements we have for our final product, which properties are important to us. First of all, we would like to highlight that the very idea of "gold" datasets (benchmarks, evaluation sets) is that they are error-free (to the extent it is possible), well-understood and well-documented, since their purpose is to provide reliable and transparent evaluation. Datasets that do not fulfill these criteria can still be useful, but they should not be marketed as gold.

Documentation is particularly important. A drawback of (Super)GLUE is that they have not always thoroughly documented changes to the original datasets. We will also undoubtedly have to convert and change some existing datasets, and we plan to document all the changes we make (and document their initial state, too, which in some cases might be problematic, if relevant documentation is lost or incomplete). Our time estimate in table 1 takes that into account. For datasets that are large enough a split into train, dev and test should be provided.

We envision the final product of this project as a page on the Språkbanken website where all the resources (and the relevant documentation and information) are gathered. Setting up a shared-task platform, like the one (Super)GLUE is using, is beyond the scope of this project. Such a platform would also enable keeping part of test data hidden, which we currently leave for the future.

It is useful to estimate how well human beings perform on the benchmarks. This has been done both for GLUE (Nangia and Bowman, 2019) and SuperGLUE (Wang et al., 2019) using hired non-expert annotators. Note that human baselines are estimated using small randomly selected samples, not the whole datasets. These estimates suggest that for the GLUE tasks machines perform on average better than humans, while SuperGLUE still provides some headroom for further progress. The results, however, vary across datasets, models and training modes (humans, for instance, are typically much better at tasks where only small training sets are available). Human baselines are beyond the scope of this project, but can be implemented in the future, either as part of a larger effort or a project on its own.

3 Recommendations for SwedishGLUE

Within the SwedishGLUE project, we recommend creating new Swedish datasets for inference/entailment, word sense disambiguation, semantic similarity, lexical relations, reading comprehension, and sentiment analysis. These datasets are of high importance and at the same time relatively straightforward to create, in some cases by extraction from already available data. In addition, we rec-

Task type	Recommendation	Time	Size	Contact
Coreference (sec. 4.1)	(Semi-)manual translation (1) SweWinograd (2) SweWinogender	 0.5 0.5	 150 sent 120 sent	Yvonne
Inference/ Entailment (sec. 4.2)	(3) SweFraCaS-2: Adapt (4) FAQ-entailment: New dataset	0.5 2	346 problems	Sasha
Word sense disambiguation (sec. 4.3)	(5) SweWiC: Extract data from Saldo, Saldo: Exempel, SweFN and Eukalyptus	1	At least 300 lemmas	Gerlof
Sentence-level semantic similarity (sec. 4.4)	(6) Paraphrase: Manual correction of an autotranslated sample	3	at least 100-200 sentence pairs	Yvonne
Lexical semantics (sec. 4.5)	(7) LexRel: Use Saldo, SweFN, Swesaurus and Synlex for dataset of relatedness, synonymy and hyper/hyponymy (8) Högskoleprovet ord	1 0.5	40+ K tokens	Dana
Sentiment analysis (sec. 4.6)	(9) Sentiments: Extract data from ABSABank	1	up to 1.5M tokens	Sasha
Linguistic acceptability (sec. 4.7)	Leave for later	-	-	-
Lexical semantic change (sec. 4.8)	(10) LexSemChange: Re-use the SemEval 2020 dataset	0	31 word pairs	Sasha
Diagnostics (sec. 4.9)	(11) Semi-manual translation	2	1,100 sent	Felix
Reading comprehension (sec. 4.10)	(12) SweSquad: semi-manually construct equivalent	1	at least 100-200 questions	Yvonne

Table 1: Recommendation for datasets to create within the SwedishGLUE project. Time estimates are given as person months.

ommend translating the English datasets for diagnostics and coreference, which may also be used for some diagnostics and bias detection. Finally, we will use an already available evaluation set for lexical semantic change.

An overview of our recommendation is found in Table 1. In section 4 we provide a detailed overview of all tasks and the rationales for our decisions. Overall, our selection is based on what we consider to be doable within the current project. While more datasets are needed for the future, we consider this to be an appropriate and broad-coverage start for a Swedish evaluation set.

Over all, we recommend the following eleven work packages where Swedish data are created or adapted for evaluation purposes. (The numbering refers to the numbers in table 1; note that their order is arbitrary.) All datasets will be created iteratively, where a first (or preliminary) version is released to the larger project group for testing and approval. For each work package we give an estimate of when the work will be carried out, as well as when a first version will be available to project members.

1. SweWinograd: October to January, review December
2. SweWinogender: October to January, review December
3. SweFraCaS-2: December, review January
4. FAQ-entailment: January to May, review February
5. SweWiC: January to March, review February
6. Paraphrase: January to May, review March
7. LexRel: February to April, review April
8. Högskoleprovet ord: October to December, review December
9. Sentiment: December to February, review January
10. LexSemChange: review December
11. SweDiagnostics: October to May, review January
12. SweSquad: January to May, review April

4 Overview of tasks

In this section we will shortly discuss the GLUE/SuperGLUE tasks and datasets, and discuss alternatives that could be used for a Swedish evaluation set. We broadly group the tasks, although they are generally difficult to categorize, because some tasks fit in several groupings.[3] The main reason for still categorizing tasks is that not all GLUE/SuperGLUE tasks will be addressed during the initial development round (i.e the SwedishGLUE project). In addition, some tasks will change compared to English, due to the availability of, or lack of, datasets for Swedish.

4.1 Coreference

The two SuperGLUE tasks *The Winograd Schema Challenge* (which is the same as *Winograd NLI* in GLUE) and *Winogender Schema Diagnostics* deal with

[3] This is also visible in the GLUE/SuperGLUE papers, where the first one categorizes tasks, while the second one omits such a grouping.

coreference resolution. The latter is considered a diagnostic in SuperGLUE, also exposing bias.

The Winograd schema dataset[4] consists of 150 sentences in two variants, which contain a pronoun. The antecedent of the pronoun differs between the two variants. In the sentence 'The city councilmen refused the demonstrators a permit because they [feared/advocated] violence' the pronoun *they* points to *the city councilmen* if the sentence contains the word *feared*, and to *the demonstrators* if the sentence contains *advocated*.

The Winogender schema dataset[5] consists of sentence templates for 60 occupations, where each occupation has two different sentences tied to it. Each sentence contains the occupation, a participant (or 'someone'), and a pronoun (female, male, or neutral). The antecedent of the pronoun can be either the holder of the occupation, or the participant.

- The nurse notified the patient that ...
  - her shift would be ending in an hour.
  - his shift would be ending in an hour.
  - their shift would be ending in an hour.
- The nurse notified the patient that ...
  - her blood would be drawn in an hour.
  - his blood would be drawn in an hour.
  - their blood would be drawn in an hour.

The tasks are considered generally easy for humans, but less so for machines (humans achieve an accuracy of 100 on WSC; while the best system's performance is 93.8), as they require background knowledge, and may expose bias (humans on AX-g: gender parity 99.3, accuracy=99.7; best system: gender parity 98.3, accuracy 99.2). These datasets are therefore useful for both the more specific question of coreference resolution, but also for detecting bias, and as a general diagnostic (see section 4.9).

As far as we know, there is no similar dataset for Swedish available. We therefore propose to (semi-)automatically translate the schemas, with manual post-processing. Considering the task and the format of the sentences, semantic differences should not be too big of a problem. We estimate at most one month of work for translating both schema sets, which would result in high-quality test sets.

4.2 Inference/entailment

Tasks dealing with inference or entailment in many respects are also about reading comprehension. Several tasks, however, strictly deal with logical entailment.

For the two MultiNLI tasks in GLUE a corpus annotated with textual entailment is used. The annotation consists of one of three labels (entailment, contradiction, or neutral, i.e neither) for sentence pairs. The texts are taken from a range of genres, and the two tasks, labelled matched and mismatched,

[4] `https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`

[5] `https://github.com/rudinger/winogender-schemas`

are evaluated in-domain and cross-domain, respectively. The corpus contains more than 400K sentence pairs. This is similar to the GLUE/SuperGLUE task Recognizing Textual Entailment (RTE), where, given a short text and a hypothesis, the task is to determine whether the text entails the hypothesis. The labels are binary; entailment and no-entailment. The RTE contains just under 6K instances, taken from news text and Wikipedia.

A broader notion of entailment is visible in the SuperGLUE tasks based on CommitmentBank (CB) and Choice of Plausible Alternatives (COPA). CB contains 1200 instances, taken from news, fiction, and dialogue, annotated with the author's degree of commitment to the truth of a clause (true, false, uncertain). Only a part of the CB data is used for the SuperGLUE task, where inter-annotator agreement was above 80%. The COPA task requires the system to determine the cause or effect of a sentence, given two alternatives. The manually annotated data of 1000 questions is extracted from blogs and a photography-related encyclopedia.

A possible alternative for Swedish would be to automatically translate parts of the data and manually curate it to make sure it has sufficiently high quality. Another would be to scrape FAQs from the websites of authorities to get question-answer pairs. To create a high-quality dataset, this would include filtering out very long answers, and scrambling questions and answers to get negative examples, as well as additional manual curation.

As a first step, however, we propose using the Swedish version of The FraCaS textual inference problem set[6] (Cooper et al., 1996; Ljunglöf and Siverbo, 2012), which consists of more than 300 problems, each with one or more statements and one yes/no-question. We estimate a small effort for data conversion and translation of some sentences, according to Ljunglöf and Siverbo (2012).

4.3 Word sense disambiguation

SuperGLUE includes Words in Context (WiC), where the task is, given two text snippets and a polysemous word that appears in both sentences (not necessarily in the same form), to determine whether the word is used with the same sense in both sentences.

It is sometimes assumed that a plateau has been reached for word-level tasks and they thus are not really worthy of attention. That, however, is not true, at least not when tasks have to do with meaning. The human baseline for WiC is an accuracy of 80.0, and the best system performs at 76.5.

The SwedishGLUE application lists WiC as a priority (though not as a first-level priority). The relevant Swedish resources we have at our disposal are Saldo (Swedish Associative Thesaurus)[7] (Borin et al., 2013), Saldo: exempel[8], Eukalyptus[9] (Adesam et al., 2018) and SweFN (Borin et al., 2010b). Saldo lists meanings of polysemous words, while Saldo: exempel provides example

[6] `https://github.com/heatherleaf/FraCaS-treebank`

[7] https://spraakbanken.gu.se/resurser/saldo

[8] https://spraakbanken.gu.se/resurser/saldoe

[9] https://spraakbanken.gu.se/resurser/eukalyptus

sentences for some of the Saldo words. Out of 2785 words presented in Saldo: exempel 296 are polysemous. Of those, 190 have exactly two meanings, the remaining 106 have more. Eukalyptus has word-sense annotation that relies on Saldo meanings for more than 60K of its almost 100K tokens. Not all of these are of gold quality, but potential deficiencies are documented. Examples from SweFN can also potentially be used.

We suggest taking Saldo meanings and illustrating them with sentences from Saldo: exempel and Eukalyptus. A bulk of this work can be done automatically, but some manual curation will be required. We estimate the required time as one person month.

4.4 Semantic similarity

GLUE features three datasets that can be classified as dealing with semantic similarity (at the sentence level): Microsoft Research Paraphrase Corpus (MRPC), Semantic Textual Similarity Benchmark (STS-B), and Quora Question Pairs (QQP). In MRPC, the task is to determine whether two sentences are paraphrases/semantic equivalents (the human baseline F1 is 86.3, best system performance 94.5). In STS-B, the task is to compute how similar two sentences are, returning a similarity score between 0 and 5 (92.7 vs 93.2). In QQP, the task is to determine whether a pair of questions (asked on the Quora website) are semantically equivalent (59.5 vs 76.1). In MRPC, the human annotation was performed by hired experts, in STS-B, it was mostly performed by crowd-sourced annotators. For QQP, we were not able to find an explicit statement about where the annotation comes from, but it seems to be the product of the efforts of website users ("Quorans") who merge questions they deem similar. The dataset creators warn that ground-truth labels may contain some noise[10].

Isbister and Sahlgren (2020) attempt to create a Swedish benchmark by automatically translating STS-B. They do not perform a systematic manual evaluation of the translated dataset, but acknowledge a high proportion of translation errors. Nonetheless, they hypothesize that since most of those errors concern vocabulary, it is possible to use the dataset for subword- and character-based models (but not word-based models). They do not, however recommend using the dataset for training or fine-tuning models.

There is little doubt that the translated dataset does contain some useful information. Nonetheless, we know it is flawed, but do not know how exactly and to what extent, and how the translation errors affect evaluation results. Furthermore, even if the automatic translation was perfect, it is not obvious that the semantic relationship between two translated sentences in one language would always be the same as between the two original ones in another language. Note that semantic similarity in general is a nebulous task: for all GLUE benchmarks, human baselines are low (especially for QQP, see also Nangia and Bowman (2019) about that), which suggests that inter-annotator agreement is also low and/or that the task itself is poorly formalizable.

[10] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

We propose to use the translated dataset as a starting point, to extract a random sample of sentence pairs (from various genres represented in the original dataset), convert the available similarity judgments from the 0-5 scale to a simpler one (0-1 or 0-2) and then manually go through the sample, correcting the translations where necessary and checking whether the judgments make sense. We aspire to annotate 100-200 sentence pairs.

In addition, we also suggest creating a semantic-similarity benchmark at *word* level and making this task part of the lexical-relation suite, see section 4.5.

4.5 Lexical semantics

As mentioned in section 4.4, it would be valuable to have a semantic-similarity dataset at word level. The absence of word-level datasets is actually a problem for Swedish NLP, since such a benchmark would be useful, for instance, for evaluation of word embeddings (Fallgren et al., 2016) or studies of lexical semantic change[11].

For words, *semantic similarity* is usually understood either as 'synonymy' or 'relatedness' (which are related, but not synonymous notions). The datasets for these and potentially other lexical relations (for instance, (co)hypo- and hypernymy) can be constructed using the existing Swedish resources, such as: (1) SALDO (Borin et al., 2013), a lexical resource with over 130K Swedish words which contains morphological and lexical-semantic information for contemporary Swedish; (2) Swedish FrameNet (SweFN) (Borin et al., 2010a), a lexical-semantic resource that has been expanded from and constructed in line with the English Berkeley FrameNet (Fillmore et al., 2003), currently containing over 1K semantic frames with nearly 38K lexical units, and more than 8K annotated sentences that were manually annotated with semantic information; (3) Swesaurus (Borin and Forsberg, 2014), a Swedish wordnet based upon fuzzy synonym sets; (4) The People's Synonym Dictionary (Kann and Rosell, 2006), a synonym resource containing nearly 40K Swedish synonym pairs.

Most of the data can be extracted from these resources automatically, possibly with some manual curation. We estimate the time required for this step as one month.

In addition we will use högskoleprovet ordförståelse[12] (Swedish Scholastic Aptitude Test, word comprehension). The data belongs to Universitets- och högskolerådet (the Swedish council for higher education) and is freely available. To date (December 2020), there are close to 800 word tasks with five answer alternatives, where the correct answer is a synonym or hypernym to the main word or phrase. We estimate about half a month for collecting the tasks and making them available in a machine readable format.

[11] https://languagechange.org/

[12] https://www.studera.nu/hogskoleprov/infor-hogskoleprovet/ova-pa-gamla-hogskoleprov/

4.6 Sentiment analysis

Sentiment analysis is not prioritized in the SwedishGLUE application, but there are datasets at Språkbanken's disposal which make it relatively easy to create a sentiment-analysis benchmark.

GLUE uses SST-2 as a benchmark for sentiment analysis, which is a converted version of the Stanford Sentiment Treebank (Socher et al., 2013). There seems to be a very complicated relation to the original SST, which is not entirely transparent from the available documentation, but the original has less sentences, a finer-grained sentiment annotation and, most importantly, a syntactic tree for every sentence. The human baseline (97.8) for this benchmark has not been surpassed, though some systems come very close (97.5).

The Swedish resources in our possession are ABSAbank[13] (Rouces et al., 2020), SenSaldo[14] (Rouces et al., 2018) and Swedish Sentiment Lexicon[15] (Nusko et al., 2016). In Absabank, each sentiment expression is annotated as a tuple that contains the following fields: one of five possible sentiment values (as in SST), the target (what the sentiment is about), the source (who holds the sentiment), and whether the sentiment expressed is ironic. The main difference from both SSTs is that Absabank's annotation is not at sentence level, but at ”sentiment-expression” (typically a word or a phrase) level. SenSaldo and the Sentiment lexicon provide sentiment annotation at word level, but the annotation is at least in part automatic, so they are not really gold data.

This discrepancy means that we either have to formulate the task as ”find the sentiment expressions within the sentence and annotate them” or to change the ABSAbank annotation. The latter may be a non-trivial task for us, while the former may be a more difficult task for the models being tested, which is not necessarily bad, given the high performance of the existing models on the (somewhat simpler) SST-2.

Thus, we currently recommend the former solution. The final decision, however, depends to a certain extent on the quality and consistency of ABSAbank annotation (which is not entirely clear from the published paper). We propose to start the work, estimate how much can be done within a month and do it.

4.7 Linguistic acceptability

GLUE includes The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2018), which consists of binary acceptability judgments about English sentences drawn from books and journal articles on linguistic theory. We are not aware of any similar data being available for Swedish, although it is certainly possible to do the same work for Swedish that Warstadt et al. did for English (some examples may be found through the LinGO Grammar Matrix project[16]). We would, however, instead prefer to create datasets that explore linguistic accept-

[13] https://spraakbanken.gu.se/resurser/swe-absa-bank

[14] https://spraakbanken.gu.se/resurser/sensaldo

[15] https://spraakbanken.gu.se/resurser/sentimentlex

[16] http://matrix.ling.washington.edu/index.html

Coarse-Grained Categories	**Fine-Grained Categories**
Lexical Semantics	Lexical Entailment, Morphological Negation, Factivity, Symmetry/Collectivity, Redundancy, Named Entities, Quantifiers
Predicate-Argument Structure	Core Arguments, Prepositional Phrases, Ellipsis/Implicits, Anaphora/Coreference Active/Passive, Nominalization, Genitives/Partitives, Datives, Relative Clauses, Coordination Scope, Intersectivity, Restrictivity
Logic	Negation, Double Negation, Intervals/Numbers, Conjunction, Disjunction, Conditionals, Universal, Existential, Temporal, Upward Monotone, Downward Monotone, Non-Monotone
Knowledge	Common Sense, World Knowledge

Table 2: Coarse and fine-grained linguistic phenomena in GLUE diagnostics.

ability from a perspective relevant for research on language learning, language planning etc. We will currently not prioritize this task.

4.8 Lexical semantic change

Lexical semantic change has recently been attracting increasing attention from the NLP community. Since we have a gold dataset at our disposal that can be used off the shelf[17] (Tahmasebi et al., 2020), we suggest including it in the collection.

4.9 Diagnostics

In the same vein as GLUE/SuperGLUE, we intend to include a so called *diagnostic dataset* in SwedishGLUE. A diagnostic dataset, also referred to as *test suite* or *challenge set*, differs from normal evaluation data in that it is made up of specific examples, often hand-crafted, meant to assess pre-defined linguistic phenomena. As such, it does not represent a natural distribution as it occurs in written or spoken language and, thus, the raw performance of the diagnostic dataset should not be used as a metric to compare the performance of models overall. Rather, it should be used as an indicator to how well a system handles the particular linguistic phenomena in question.

[17] https://spraakbanken.gu.se/resurser/semeval2020

Defining which linguistic phenomena to cover is one of the greatest challenges in creating a diagnostic dataset. Luckily, the authors of GLUE and their predecessors of FraCaS (Cooper et al., 1996) have laid out a set of categories meant to cover a broad set of linguistic phenomena. These categories are made of up to four coarse-grained categories and more than 40 fine-grained subcategories, illustrated in table 2.

The diagnostics in (Super)GLUE are made in the context of the MultiNLI tasks defined in GLUE and the classifiers for those tasks are used to evaluate the diagnostic dataset. Thus, each entry in the GLUE Diagnostic Dataset contains a pair of sentences, where one is a conclusion and the other is the premise, a label specifying whether the relation between the sentences is a contradiction, entailment or neutral, and, importantly, the linguistic phenomena found in those sentences. The dataset consists of 1100 such sentence pairs. Since the distribution of classes is imbalanced, R3, a three-class generalization the Matthews correlation coefficient, is used for evaluation.

For the purposes of SwedishGLUE, the construction of a diagnostic dataset for Swedish will be dependent on the selection of linguistic phenomena and the choice of data. Due to the high-level nature of the linguistic phenomena and the MultiNLI task at hand, many of the sentences in the GLUE diagnostic dataset would keep the same linguistic phenomena if translated directly from English to Swedish. A few categories, such as prepositional phrases or datives, are specific to English syntax, however, due to the grammatical similarities between Swedish and English we estimate that most of these would work for Swedish as well with perhaps only slight modification.

What is important to note is that the usefulness of the diagnostic evaluation will be highly dependent on the quality and quantity of the data as well as the trained classifier used for the inference and entailment tasks. If the performance of the classifier on inference and entailment is not good in general, the diagnostic will have little to say about what linguistic phenomena it has or has not captured.

Apart from complementing the rest of the SwedishGLUE project, this dataset is also a unique contribution in that its one of the first of its kind for Swedish. Although some preliminary work has started for machine translation (Ahrenberg, 2018) and the FraCaS test suite has been semi-automatically translated to Swedish Ljunglöf and Siverbo (2012), at the time of writing, there are no other datasets that would be of the same size for Swedish.

Since the data creation process involves delicate consideration of high level linguistic phenomena, we suggest using an expert in linguistics to be involved in the post-processing of the translations. With the data creation work and the extensive documentation that would be required for the linguistic categories, we estimate the diagnostic dataset can be done in two months.

Another very recent promising approach to diagnostics is to use psycholinguistically inspired datasets (Ettinger, 2020). We suggest considering it as a future possibility, but not part of this project.

4.10 Reading comprehension

The category of reading comprehension to some extent overlaps with tasks dealing with inference/entailment (see section 4.2) and question-answering tasks. Here we will discuss the SuperGLUE tasks Multi-Sentence Reading Comprehension and Reading Comprehension with Commonsense Reasoning, as well as the SuperGLUE task BoolQ and the GLUE task Question NLI.

The dataset used for the Question NLI task originally consists of question-paragraph pairs, where one of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the corresponding question (written by an annotator). The data was converted for the task into question-answer pairs, with the labels entailment/not-entailment. The task contains 100K such pairs.

The BoolQ tasks requires that the system, given a short text passage and a question, determines whether the answer to the question is true or false. The data consist of close to 16K such triples (question, passage, answer), extracted from Wikipedia.

For the task of Reading comprehension with commonsense reasoning each example consists of a news article and a Cloze-style question about the article in which one entity is masked out. The system must predict the masked out entity from a list of possible entities in the provided passage, where the same entity may be expressed with multiple different surface forms, which are all considered correct. There are more than 120K articles taken from news texts in the evaluation set.

The Multi-sentence reading comprehension task, finally, presents a short text (paragraph) and a question about the text. There is no pre-specified amount of correct answers to each question, and the answers are not required to be a span in the text. The task contains 10K paragraph-question-answers-tuples, taken from seven different domains, such as news, fiction, and historical text.

All four of these tasks deal with answering questions based on a short text. While it would be possible to translate some of the English data into Swedish, we can assume that quality checks and corrections would take a lot of time, to achieve high-quality data. Another possibility would be to create a small dataset by extracting texts, e.g. from Swedish Wikipedia, and manually construct questions. Or we could scrape FAQs from the web, with an additional manual effort of curating the data (see also section 4.2).

We would also have like to re-use högskoleprovet läsförståelse[18] (Swedish Scholastic Aptitude Test, reading comprehension). Högskoleprovet is given twice a year to test various aspects of language skills and logical reasoning, and students can then use the result as an alternative means to school grades to apply for university studies. As of now (December 2020) there are 48 shorter texts (less than one page but with multiple paragraphs) with 2 questions each, and 46 longer texts (around one and a half page) with 4 questions each. Each question has 4 alternative answers, of which only one is correct. Generally, this makes högskoleprovet läsförståelse most similar to the BoolQ task, since the

[18] https://www.studera.nu/hogskoleprov/infor-hogskoleprovet/ova-pa-gamla-hogskoleprov/

system has to determine if an answer is true or false, while the texts are longer than for any of the GLUE/SuperGLUE reading comprehension tasks. Unfortunately, in contrast to the word task (section 4.5) each text has a different copyright owner, and so trying to get permissions is not possible within the current project.

We suggest re-using the English SQuAD data. Instead of translating the whole data set for SweSquad, we propose applying the method by Vakili (2020), who uses multilingual sentence embeddings to find answers to English SQuAD questions in Swedish Wikipedia articles. This reduces the translation work, but still requires manually selecting the Swedish answers, and translating the questions from English to Swedish. We estimate about one month of work for a smaller set of questions.

Acknowledgments

This report would not have been written without extensive discussions with our colleagues Dana Dannélls, Markus Forsberg, Gerlof Bouma, and Lars Borin.

References

Yvonne Adesam, Gerlof Bouma, Richard Johansson, Lars Borin, and Markus Forsberg. The Eukalyptus treebank of written Swedish. In *Seventh Swedish Language Technology Conference (SLTC), Stockholm, 7–9 November 2018*, 2018.

Lars Ahrenberg. A challenge set for English-Swedish machine translation. In *Proceedings of The Seventh Swedish Language Technology Conference (SLTC-2018)*, pages 27–30, Stockholm, 2018.

Lars Borin and Markus Forsberg. Swesaurus; or, the Frankenstein approach to Wordnet construction. In *Proc. GWC 2014*, pages 215–223, Tartu, 2014. GWA.

Lars Borin, Dana Dannélls, Markus Forsberg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. The past meets the present in Swedish FrameNet++. In *Proceedings EURALEX 2010*, pages 269–281, Leeuwarden/Ljouwert, 2010a. Fryske Akademy.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. The past meets the present in swedish framenet+. In *14th EURALEX International Congress*, 2010b.

Lars Borin, Markus Forsberg, and Lennart Lönngren. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211, 2013.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium, 1996.

Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020. doi: 10.1162/tacl_a_00298. URL `https://doi.org/10.1162/tacl_a_00298`.

Per Fallgren, Jesper Segeblad, and Marco Kuhlmann. Towards a standard dataset of Swedish word vectors. In *SLTC 2016 Proceedings*, Umeå, 2016. Umeå Univ.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250, 2003.

Tim Isbister and Magnus Sahlgren. Why not simply translate? a first Swedish evaluation benchmark for semantic similarity, 2020.

Viggo Kann and Magnus Rosell. Free construction of a free Swedish dictionary of synonyms. In *Proc. NODALIDA 2005*, pages 105–110, Joensuu, 2006. University of Eastern Finland.

Peter Ljunglöf and Magdalena Siverbo. A bilingual treebank for the FraCaS test suite. In *SLTC-2012, 4th Swedish Language Technology Conference, Proceedings of the Conference*, 2012.

Nikita Nangia and Samuel R. Bowman. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1449. URL `https://www.aclweb.org/anthology/P19-1449`.

Bianka Nusko, Nina Tahmasebi, and Olof Mogren. Building a sentiment lexicon for Swedish. *Linköping Electronic Conference Proceedings*, 126(006):32—-37, 2016. URL `http://www.ep.liu.se/ecp/126/006/ecp16126006.pdf`.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. SenSALDO: Creating a sentiment lexicon for Swedish. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation, 7-12 May 2018, Miyazaki (Japan)*, Miyazaki, 2018. ELRA. ISBN 979-10-95546-00-9.

Jacobo Rouces, Lars Borin, and Nina Tahmasebi. Creating an annotated corpus for aspect-based sentiment analysis in Swedish. In *Proc. DHN 2020*, pages 318–324, Aachen, 2020. CEUR-WS.org. `http://ceur-ws.org/Vol-2612/short18.pdf`.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1170`.

Nina Tahmasebi, Simon Hengchen, Dominik Schlechtweg, Barbara McGillivray, and Haim Dubossarsky. Swedish Test Data for SemEval 2020 Task 1: Unsupervised Lexical Semantic Change Detection, February 2020. URL `https://doi.org/10.5281/zenodo.3730550`.

Thomas Vakili. A method for the assisted translation of qa datasets using multilingual sentence embeddings. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2020. `http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-281826`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280, 2019.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.

GU-ISS, Forskningsrapporter från Institutionen för svenska språket, är en oregelbundet utkommande serie, som i enkel form möjliggör spridning av institutionens skriftliga produktion. Det främsta syftet med serien är att fungera som en kanal för preliminära texter som kan bearbetas vidare för en slutgiltig publicering. Varje enskild författare ansvarar för sitt bidrag.

GU-ISS, Research reports from the Department of Swedish, is an irregular report series intended as a rapid preliminary publication forum for research results which may later be published in fuller form elsewhere. The sole responsibility for the content and form of each text rests with its author.

GÖTEBORGS
UNIVERSITET