university of groningen / faculty of mathematics and natural sciences / artificial intelligence

Recognising Textual Entailment across English and Dutch for humans and the ECNU-system

Bachelor's Project Thesis

Joost Doornkamp, s2550156, j.doornkamp@student.rug.nl,
Supervisors: Dr Jennifer K. Spenader & Hessel Haagsma

Abstract: In the Recognising Textual Entailment (RTE) task the goal is to predict whether one sentence logically entails another. Given a translated RTE data set, we investigate whether the ECNU Support Vector Machine (SVM) for RTE performs as well on Dutch text as it does on English. For this we recreate the ECNU system for English and Dutch by extracting features from th sentence pairs and compare the performance of both versions as well as the values of individual features in both languages, to see whether the representation of sentences is constructed similarly. We also investigated the validity of the translated set in an annotation study. We conclude that there are likely no significant differences between English and Dutch in the context of RTE and that the translation of RTE data sets into new languages is a promising method for kickstarting the development of RTE systems for those languages.

1 Introduction

In the RTE task, systems predict whether one sentence (called *text*) has another sentence (called *hypothesis*) as a logical conclusion, i.e. whether the text entails the hypothesis. For example, given the following text-hypothesis pair (t-h pair):

t: "*Debbie Inker is a CPA who has lived in Israel since 2002.*"

h: "*Debbie Inker is a citizen of Israel.*"

a RTE system has to predict NO, as the hypothesis does not entail the text (living in Israel does not mean Debbie Inker is a citizen there). RTE systems are often supervised machine learning applications where the correct entailment relation for sentence pairs has been annotated by humans. These data sets are typically in English, but recently the English data set for the Third PASCAL RTE challenge (Giampiccolo et al., 2007) was translated to Dutch as part of the Parallel Meaning Bank (Bos et al., 2016). It is the first such data set translated to Dutch.

In this investigation we aim to explore whether the translated system will perform similarly on the translated data set in comparison with the English system with the original data set. The goal is to both validate the translation of the data set (which we will do as a small separate study) and the idea that RTE systems can be adapted from English to Dutch, and maybe other languages, rather than having to develop new systems for each language.

To test this we recreate the SVM version of the ECNU-system (Zhao et al., 2014), which had the second best score on the Textual Entailment subtask of SemEval-2014 task 1 (Marelli et al., 2014), for both English and Dutch. These systems are then applied to the (translated) data set and compared.

The goal of translation is of course to carry over the same semantic information to another language, but subtle grammatical, lexical or other differences between languages may make this impossible. Humans are typically able to compensate for this by using different words or more elaborate descriptions, but such differences may prove a new problem altogether for Cross-Lingual Textual Entailment (CLTE). This study will evaluate whether these differences prove an obstacle for translated RTE.

In the next section we will discuss some more background of the ECNU system and RTE. After that we will discuss the data set, the data set validation and the constructions of the systems and their performance. In the last section the two parts will be combined into the final conclusion: the translations are good and translating the ECNU system does not seem to have a negative influence on its performance.

2 Background

2.1 The task

The Recognizing Textual Entailment (RTE) task is a popular NLP task, which started in 2005 with the first PASCAL RTE challenge (Dagan et al., 2006) and featured in the SemEval workshops up to 2014 (Marelli et al., 2014). RTE systems are used for question answering systems, inference systems (Magnini, 2015) and even student response analysis systems, where the goal is to automate assessment of student responses to questions (Dzikovska et al., 2016). The task has also been done cross-lingually for e.g. German, French and Italian (Negri et al., 2012).

To explain what is important in RTE, we are going to compare it with the task's metaphorical brother, Semantic Text Similarity (STS). In STS, the goal is to assess whether two sentences have a similar meaning, which can be assessed by comparing representations of semantic information in the sentences (Agirre et al., 2012). The better the representation, the better the meaning of a sentence has been modelled and thus the better the meanings can be compared. In RTE the relation between two sentences needs to be compared as well as their individual meaning.

RTE is no longer featured in SemEval while STS is, most likely because STS is simpler and focuses more on accurately representing semantic information. However, is problematic in that 'true' semantic similarity is hard to measure. In SemEval the 'true' similarity is annotated by humans using a ranking system (e.g. 1 is 'not similar' and 5 is 'very similar'/'the same'). Humans often do not agree on scores. In RTE this agreement is more straightforward, a sentence entails another or it does not. Because it is more clear-cut it is easier for human annotators to assess, and simpler for machines to predict.

RTE still comes down to having an accurate representation of the semantic information of sentences though. The big difference is that for STS it often suffices to represent the semantic information of both sentences separately and compare those afterwards, while for RTE the interaction between the representations is important. Thus, RTE systems benefit from using only one representation that relates to both sentences as well as their relation. For example, a feature-based approach may have features relating only to the text, features relating only to the hypothesis and features relating to both at the same time. This is the case for the ECNU system. Using one, holistic representation is not unique for RTE though: the ECNU system also works for STS using the same representations.

2.2 SemEval-2014

The ECNU system is an example of a Compositional Distributional Semantic Model (CDSM), which look not only at the word level and the distribution of words across corpora but also represent semantic information of phrases and sentences, of which SemEval-2014 was the first development that made benchmarks to test such systems. In SemEval-2014 21 teams participated in the same task using the same data set (the SICK data set), thus allowing for the systems to be compared objectively. This was not the first time RTE featured in SemEval, but SemEval-2014 focused specifically on CDSMs. The ECNU system is compositional only at sentence level, it uses no phrasal information.

Compositional models are quite a development since the systems that were built for the third PASCAL RTE challenge in 2007, where most systems were either using lexicon-based (using resources like WordNet and DIRT) and syntactic features or transformations on dependency structures. In CDSMs we see the combination of both as well as entirely new methods such as the use of denotational similarities (Young et al., 2014). Machine Learning has both been applied more and the methods have significantly improved, now also using methods like neural networks to form language models that represent the semantic information.

Given these developments it comes as no surprise that the best result for the PASCAL challenge was only 67.0% accuracy, while in SemEval-2014 it reached 84.6%. Note though that these workshops have very different data sets, and thus these results are not directly comparable. It merely illustrates that overall performance on RTE has improved.

In SemEval-2014, ECNU scored 83.6%, coming in second, right behind Illinois-LH (Lai and Hockenmaier, 2014) (which got the 84,8%). For this study we chose to use the ECNU system over Illinois-LH because the implementation was expected to be quicker and simpler. Additionally, the ECNU system defined its features in clear, modular groups, which allowed us to leave some groups out for this preliminary study.

So can we expect the recreated ECNU system to reach similar accuracy for this study? First of all, recall that the data set used here and the SICK data set are different, so there is no real reason to assume that we can reach the same accuracy. We also do not use all features that are used in the ECNU system, so accuracy will likely be lower.

Additionally, Marelli et al. (2014) has expressed concerns that some techniques used by participants in SemEval-2014 may have been too *ad-hoc*, using not representations of the semantic information in the sentence pairs but rather specific properties of the data set. However, if they are present they likely are related to the way the data set was constructed, and because the PASCAL data set has very different sources this will likely not translate to the recreated system used in this study. Thus, if the SICK data set used in SemEval-2014 did have some structural properties relating to the entailment relation and the ECNU system exploits these, it will negatively affect performance.

Marelli et al. (2014) fails to mention exactly what properties would be exploitable this way. As said before, the data sets have the same format for its sentence pairs, but it may be the case that systems use information such as question order or keywords to detect a certain type of sentence pair and use that information for its predictions. Note that these are merely examples of non-semantic information that could be used, we have no reason to assume this is the case.

Also note that in this study we do not actually use the original ECNU system but rather recreate it, following the documentation in Zhao et al. (2014) closely. This recreation consists of taking their way of representing the semantic information (via features, see Section 5.1.2). This way, if any non-semantic methods existed in the original system, it will likely not be present in the recreated system, or we will have noted it during recreation.

As will be discussed in Section 6.3, we ultimately conclude that we can not exclude this being an effect, as not all features are used for this study and thus the semantic representation is not the same. The results of the original system and the recreated one will not be not comparable. This will require further research.

For a more detailed summary of the developments in RTE, refer to Magnini (2015), a book written by the first author of the first PASCAL RTE challenge in which he reflects on the last decade of RTE research.

2.3 The ECNU System

The ECNU system (Zhao et al., 2014) uses 7 categories of features with a total of 72 features in various classifiers, such as k-nearest neighbors (kNN), Gradient Boosting (GB), Random Forest (RF) and Support Vector Machine (SVM). In this study only SVM was used as Machine Learning method and only 32 distributional features are used however, so the recreated systems actually are not CDSMs. We will recommend further research to implement the other 40 features, including the compositional ones (see Section 6.3).

The fact that the ECNU system was made for the SemEval-2014 task means that the results of this study (in which we use the data set of the third PASCAL RTE challenge) will not be completely comparable with that of SemEval. The sentence pairs in the two data sets have the same format, but the goal of the RTE task in SemEval-2014 was to predict 'entailment'/'contradiction'/'neutral'. Neutral in this case would mean that no entailment was present but also no contradiction. In the PASCAL challenge only 'entailment' and 'no entailment' was predicted.

We think this 'downgrade' will not pose a problem for the ECNU system, because whenever it would have predicted 'contradiction' or 'neutral', it should now simply predict 'no entailment'. If the SVM was able to learn this three-value classification it must be able to learn this two-value classification as well.

To illustrate, imagine a version of the ECNU system that was trained on three-valued classification. This system could easily be converted to two-value classification by mapping each instance it classifies as 'contradiction' or 'neutral' to 'no entailment'. An SVM can learn such a one-to-one mapping, and because we know it can also learn the three-value classification we know that it can learn the two-value classification.

Additionally, the SICK data set used in SemEval-2014 had 10,000 sentence pairs while the data set of the PASCAL challenge, and consequentially the translated data set, had only 1,600 pairs. A larger data set is of course preferable, but keep in mind that this is a preliminary study. As will be discussed in Section 6.3, the positive result of the translated data set validation will prove that RTE data sets are translatable (between English-Dutch). This will hopefully lead to more data sets being translated, and similar studies may be carried out on larger data sets.

3 Data Set

The original, English data set is the data set used in the third PASCAL RTE Challenge (Giampiccolo et al., 2007). This data set was translated to Dutch as part of the Parallel Meaning Bank project (Bos et al., 2016). Both the English and Dutch versions are used for this study. At the time of creation no evaluation of translation quality was done, so before it was used for the recreated system a validation test was done to see whether it was reasonable to assume that the translations are correct, i.e. whether the meaning of the sentences, and their entailment relation, have stayed the same (see Section 4).

The data set was already divided into a training and a test set, where both sets consist of 800 sentence pairs. The sentences are also categorised as short or long, where long means the character string exceeds 270 bytes. Lastly the sentence pairs are divided by the task for which the pairs were originally generated:

- **IE**: The Information Extraction task focused on making sentence pairs (text-hypothesis) from various other tasks. In these cases IE systems provided the hypotheses given some texts. The annotation of entailment relations was then added by humans. Some pairs were also manually generated by humans.

- **IR**: In Information Retrieval a propositional query was was taken from TREC (Text Retrieval Conferences) and CLEF (Cross-Language Evaluation Forum) data sets as a hypothesis, and associated documents were retrieved from various search engines. The entailment relation between the text, taken from the document, and the hypothesis was then annotated by humans.

- **QA**: In the Question Answering task annotators generated hypotheses from questions from the TREC and CLEF data sets, amongst others, by combining the question and its answer - as given by various QA systems - into an affirmative sentence. The texts were generated from the systems' answers (the QA systems answered with a relevant paragraph from the Internet, without extracting key information from it).

- **SUM**: In the Summarization task the sentence pairs were based on summarization systems, which combine passages from semantically related documents into clusters. The hypotheses were taken from the overlap in the cluster (given by some system directly as a cluster name or otherwise selected by annotators), while the texts were selected by the annotators from the cluster.

During creation of this data set care was taken that, regardless of the original task, the sentence pairs had the same format. Thus, these origins should have no effect on the format of the sentence pairs and thus on annotation by humans nor the performance of the system. This is checked in Section 5.2.

In the original study the entailment value for all sentences was annotated by three human annotators and pairs where the annotators disagreed were removed. Several other pairs were removed due to being controversial, considered too difficult or because a pair that was too similar was already in the set (Giampiccolo et al., 2007).

4 Translated data set validation

4.1 Methods

The validation of the translated data set was done by having human annotators assess entailment for a sample of the translated data and checking agreement with the original (English) annotations in the RTE data set.

32 Dutch text/hypothesis pairs were randomly selected from the data set, with equal distribution over which set it came from (test/train), length class (short/long), correct answer (yes/no) and task

(IE/IR/QA/SUM). This sample was ordered randomly[1]. The pairs were presented online using SurveyGizmo[2] to a set of 23 subjects. The subjects were recruited using an email to associates with a request to forward the mail to as many people as possible. At the start of the survey subjects' age was collected, along with whether Dutch was their native tongue and whether they spoke any other languages in kindergarten. Most subjects were between 19 and 26 years old, but five subjects were between 45 and 56. Participant gender was not recorded.

The sample size was set at 32 sentence pairs to maximise the amount of questions without fatigue effects occurring. Subjects were also recommended to take as much time as they needed and take breaks if necessary even though the survey generally took only about 15 to 20 minutes. Before the survey started there was a brief instruction on entailment, including an example of a non-entailing sentence pair. In this sentence pair the two sentences were entirely unrelated but the hypothesis was a well-known fact; this was to make clear that the hypothesis has to follow from the text, regardless of its absolute (real world) truth value.

For each pair, the subject was presented with the text at the top with the hypothesis below it, marked 'entailment'. Below that was the question whether the entailment was correct. Subjects could choose either yes or no, there was no neutral option. Only one question was presented at a time to minimise distraction.[3]

4.2 Results

In Figure 4.1 the results of the data evaluation are shown. We see that for most questions subjects agreed with the annotated values, as for 26 out of 32 questions more than 20 out of 23 subjects agreed with the entailment value in the data set. However, there are a handful of exceptions. Most noticeable are questions 13, 17 and 20, where more than half of the subjects disagreed with the annotated entailment value, suggesting that the alternative may be preferred. These questions, the set these question come from and the corresponding identification number and the text and hypothesis can be found in Table 4.1.

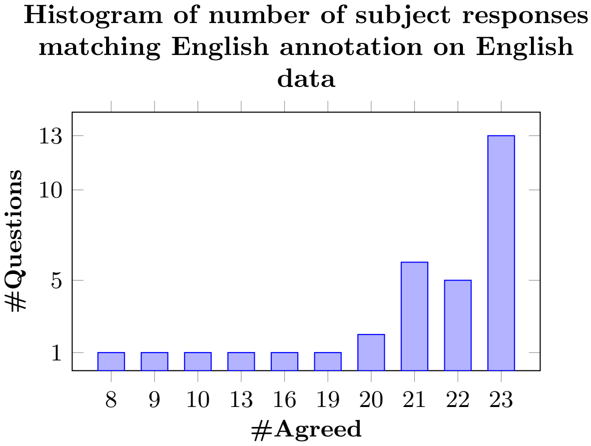

Figure 4.1: The number of questions where a certain number of subjects agreed with the annotated value in the English data set.

Discussion with participants after the survey revealed that the first sentence of question 13 was perceived to be very confusing. This does not seem to be a result of the translation as the translated sentence has a very similar structure. This structure is likely the cause of the high level of disagreement (with original annotation but also with other subjects), because there is ambiguity as to what '*founder*' relates to ('*Kalido Technical Advisory board*' or '*Textra Group*').

Recall that the data set was originally annotated by three annotators, and pairs that annotators disagreed on were removed. It is unknown whether differences between the original annotators and the subjects in this survey (e.g. in age or education level) can explain why the Dutch annotators seem to find this sentence structure more difficult. Even if it can though, it must be noted that the sample of three annotators is relatively small, so general disagreement regarding certain sentence pairs may have been completely missed in the original annotation.

The first sentence of question 17 contains what is likely a translation error: the English 'touched down' translates directly to 'landde', but during translation this constituent has moved from after the comma to in front of it. This makes the subor-

[1] Random sequence determined by `https://www.random.org/sequences/`

[2] `https://www.surveygizmo.com/`

[3] The survey can be found at `http://www.surveygizmo.com/s3/2688730/Translated-RTE-Pilot` (Dutch)

Q#	ID	Set	Text	Hypothesis	#A
13	23	Train	*De leden van de Kalido Technical Advisory Board omvatten Boris Evelson, oprichter en managing partner, Textra Group, Inc, en Bill Inmon, voorzitter, Inmon Data Systems.* The Kalido Technical Advisory Board members include Boris Evelson, founder and managing partner, Textra Group, Inc., and Bill Inmon, president, Inmon Data Systems.	*Boris Evelson richtte de Kalido Technical Advisory Board op.* Boris Evelson founded the Kalido Technical Advisory Board.	8
17	777	Train	*Het Hercules-transportvliegtuig dat direct hierheen vloog vanaf de eerste ronde van de reis in Pakistan landde, en toen was het slechts een vlotte wandeling van 100 meter naar de handdruk.* The Hercules transporter plane which flew straight here from the first round of the trip in Pakistan, touched down and it was just a brisk 100m stroll to the handshakes.	*Het Hercules Transporter-vliegtuig maakte een vlucht naar Pakistan.* The Hercules transporter plane made a flight to Pakistan.	9
20	431	Test	*In een interview met de Sci Fi Wire zei Robert Shaye, co-voorzitter van New Line en executive producer van "The Lord of the Rings," dat New Line in de toekomst geen enkele film, waaronder "The Hobbit" wil doen met meneer Jackson.* In an interview with the Sci Fi Wire, Robert Shaye, co-chairman of New Line and executive producer of "The Lord of the Rings," said that New Line does not want to do any films, including "The Hobbit," with Mr Jackson in the future.	*"The Lord of the Rings" werd geregisseerd door Peter Jackson.* "The Lord of the Rings" was directed by Peter Jackson.	10

Table 4.1: The three questions with the highest disagreement. Q# is the question number in the survey, ID the number of the sentence pair in the part of the data set as indicated by Set (Train/Test). #A is the number of people who agreed with the entailment relation in the data set.

dinate clause structure of the Dutch sentence ungrammatical and unclear.

However, this may not be the only cause of the high level of disagreement. The first sentence clearly mentions a flight to *here* (the exact location is irrelevant) and a trip *in* Pakistan, but never a flight *to* Pakistan. It is implicit that the plane originated from *here* and thus a trip to Pakistan is required to make a trip back, but instructions given at the start of the survey stressed that the second sentence must follow strictly from only the first sentence, no other information was to be used. So, the disagreement may also be explained through different interpretations of this instruction and the sentence. Some may think that the trip from Pakistan does follow from making a trip to Pakistan, while others may think that to be external information and thus they should answer "no".

In question 20 the annotated entailment relation appears solid (nothing is said about Jackson directing anything), but the fact in the hypothesis describes something that is relatively well known in the subject pool to be true may lead participants to choose entailment. Informal post-survey discussions revealed that some subjects found difficulty in using only the first sentence as information source. One reported "knowing that the right answer was no, but still wanting to press yes". Recall that an example was given in the instructions before the survey to address exactly this, so it is worrisome that it may still have an effect. It may also be the case that these instructions were forgotten by question 20, which would mean fatigue effects also played a role in the disagreement.

The above explanation suggests that the method of acquiring subjects (email and mouth-to-mouth, including social media) may have caused a bias in the subject pool, particularly focusing on the more highly educated, younger individuals. The mean reported age is approximately 29, but the distribution shows that 18 out of 23 subjects are between 19 and 26, and the other five are between 45 and 56 years old. This likely corresponds to the researcher's generation and that of his parents, resp. It seems probable that subjects have come from a relatively close social circle, which is why the sample may be biased. Because of the survey's anonymity other attributes cannot be traced.

The above examples explain why such a strong difference can exist between questions. There is a strong distinction between the (majority of the) questions, where the subjects agreed with the original annotation, and the questions where not only agreement with the annotated value drops but internal agreement also drops dramatically. The mean kappa score of agreement between subjects is 0.8547, but the kappa for only questions 13, 17 and 20 the kappa drops to 0.5046. The fact that it is not lower supports the idea that when agreement is low it is not because the annotated relation is wrong, but rather because the actual entailment relation is unclear, for various reasons. It is never the case that a sentence pair is annotated by English annotators as entailing while all Dutch annotators say that it is not; the English annotations are never *wrong*, the entailment relations are just ambiguous.

In Giampiccolo et al. (2007), the original annotation experiment that took place after creation of the English data set, an average kappa score of 0.75 was achieved. After that the pairs on which annotators disagreed were removed, so in the English data set we can assume that all English annotators agreed with each other.

Our kappa score is well above this 0.75, it is even a category higher in the performance measures established by Landis and Koch (1977), acquiring the label 'almost perfect'. This label is essentially arbitrary, but it indicates that the agreement is well above standards and the odds of the observed agreement occurring purely by chance is small. However, it is definitely not the case that Dutch annotators agreed on all sentence pairs as the original annotation would suggest. This likely has to do with the fact that only three annotators were used, the chance that three annotators agree on an ambiguous entailment relation is simply a lot higher than that 23 annotators do.

Nonetheless the kappa score is fairly high, largely shifted by a few ambiguous questions. These ambiguous questions may be changed or removed for later research, but even with them included in the data set the kappa score is at a satisfactory level. We conclude that the Dutch subjects mostly agreed with the entailment relation annotated in the data set, thus we find the translated data set suitable to use for the rest of the research.

5 Dutch & English RTE Systems

5.1 Methods

We used a SVM as learning algorithm, as it had the best performance as reported by Zhao et. al. (83.46%, which was 0.3% higher than RF), and the first 32 of the features used by the ECNU system (see Section 5.1.2). The English and Dutch versions of the system were developed in parallel.

Note that the SVM algorithm used for both systems is the same, but the way features are extracted differs for the two systems. These differences will be discussed in more detail in Section 5.1.2.

Reproduction of the ECNU system was kept as close to the original as possible, but the external resources used by the system were often only available for English and thus had to be substituted by Dutch alternatives.

5.1.1 Preprocessing

The ECNU system used a phase of prepossessing to normalise the data and generally make sentences more similar, in order to improve the representativeness of the semantic features. This preprocessing phase has three parts:

- **Contraction Normalisation**: the substitution of contractions such as "*hasn't*" to "*has not*".

- **Lemmatisation**: the reduction of every word to its base form. For example "*went*" would be lemmatized to "*go*".

- **Synonym Normalisation**: the replacement of a word by a synonym. If a synonym of a word in the text is present in the hypothesis, replace it so that the same words appear in both sentences.

Contraction normalisation in English was done using a lookup table. Zhao et al. (2014) did not provide a list of the contractions that were used, so an original one was used[4]. Contraction normalisation was not done for the Dutch data as Dutch does not feature contractions.

The ECNU system used the Natural Language Toolkit (NLTK)[5] to find a word's lemma in WordNet. For both the English and Dutch systems we used SnowballStemmer.[6] We chose to use this module - which stems words rather than lemmatising them - because this module was available for both languages and will likely operate the same way, reducing the difference the external resource will make regarding performance. Note that for its purpose, removing inflection from the words, stemming and lemmatisation does not make a difference.

The ECNU system used WordNet to find synonyms for synonym normalisation. The English system also uses WordNet, but the Dutch system uses Open Dutch Wordnet (Postma et al., 2016). This is a Dutch variation on WordNet, based on Cornetto (Vossen et al., 2013).

5.1.2 Features

Three categories of features were implemented following the instructions in (Zhao et al., 2014): length features (*len*, 16 total features), surface text similarity (*st* 10 total features) and semantic similarity (*ss*, 6 total features). This resulted in a total of 32 features. The algorithm for computing them is independent of language can be used to calculate them from both the Dutch and the English data sets. This algorithm was implemented in Matlab and is available online.[7]

The first half of the length features (*len*) features is based on unique word count for each of the sentences and the overlap/difference. The features are shown in Table 5.1. In this table, A stands for the set of words in the text and B for the set of words in the hypothesis.

Name	Formula
numUniqueWordsA	$\|A\|$
numUniqueWordsB	$\|B\|$
sizeDiffA	$\|A - B\|$
sizeDiffB	$\|B - A\|$
sizeUnion	$\|A \cup B\|$
sizeIntersect	$\|A \cap B\|$
normDiffA	$\frac{\|A\|-\|B\|}{\|B\|}$
normDiffB	$\frac{\|B\|-\|A\|}{\|A\|}$

Table 5.1: The first eight length features.

The second half of the length features are based on POS-tags. POS-tagging for the English system was done using the Stanford POS Tagger[8]. For Dutch, Alpino was used (Bouma et al., 2001). Given the POS tags from these systems, the set A and B were this time constructed from only the set of words in the sentences that were either nouns, verbs, adjectives or adverbs. This gives us eight sets, namely the set of all nouns in the text (A_{noun}) and all nouns in the hypothesis (B_{noun}), all verbs in the text (A_{verb}) and all verbs in the hypothesis (B_{verb}), etc. The eight features extracted from this are $|A - B|$ and $|B - A|$ for each pairs of these sets:

Name	Formula
sizeDiffNounA	$\|A_{noun} - B_{noun}\|$
sizeDiffNounB	$\|B_{noun} - A_{noun}\|$
sizeDiffVerbA	$\|A_{verb} - B_{verb}\|$
sizeDiffVerbB	$\|B_{verb} - A_{verb}\|$
sizeDiffAdjA	$\|A_{adj} - B_{adj}\|$
sizeDiffAdjB	$\|B_{adj} - A_{adj}\|$
sizeDiffAdvA	$\|A_{adv} - B_{adv}\|$
sizeDiffAdvB	$\|B_{adv} - A_{adv}\|$

Table 5.2: The last eight length features.

[4] Available on https://github.com/Superkebabbie/Translated-RTE

[5] http://nltk.org

[6] http://www.nltk.org/api/nltk.stem.html

[7] https://github.com/Superkebabbie/Translated-RTE

[8] http://nlp.stanford.edu/software/tagger.shtml

The first four Surface Text Similarity features (*st*) are again based on the sets of words in the two sentences, as shown in Table 5.3.

Name	**Formula**	**Name**	**Formula**
jaccard	$\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	overlapA	$\frac{A\cap B}{\lvert A\rvert}$
dice	$2\cdot\frac{\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$	overlapB	$\frac{A\cap B}{\lvert B\rvert}$

Table 5.3: The first four surface text features.

The other 6 features are based on the tf*idf vector representations of the text and hypothesis: where $\vec{x}$ and $\vec{y}$ are the vector representations of the text and hypothesis, resp. The last three surface text features are the Pearson, spearman and Kendall's tau (named kendall in the system) correlation coefficients of the two vector representations.

Name	**Formula**
cosine	$\frac{\vec{x}\cdot\vec{y}}{\lvert\lvert\vec{x}\rvert\rvert\cdot\lvert\lvert\vec{y}}$
manhattan	$\sum_{i=1}^{n}\lvert\vec{x}_i-\vec{y}_i\rvert$
euclidian	$\sqrt{\sum_{i=1}^{n}(\vec{x}_i-\vec{y}_i)^2}$

Table 5.4: The next three surface text features.

The Semantic Similarity features (*ss*) used Weighted Matrix Factorisation (WTMF) (Guo and Diab, 2012). This method was more recently updated to Orthogonal Matrix Factorisation (ORMF) (Guo et al., 2014), which was used to get vector representations of all sentences. The ORMF representation was reported to model the contextual meaning, especially in short texts, very well. Given these new vector representations, the six features extracted are the cosine, Manhattan, Euclidian, Pearson, Spearman and Kendall's tau vector distances. These features will be "ormf cosine", "ormf manhattan", etc. in the rest of this study.

5.1.3 SVM

After extracting the features for both languages the SVM was trained on the corresponding training set (English/Dutch) and then used to predict entailment relations for the corresponding test set. The data set was already split into a training and a test set (both 800 sentence pairs); this structure was also carried over to the translated data set. Each prediction was compared with the annotated relation in the data and performance was measured as accuracy.

5.2 Results

5.2.1 Performance

Table 5.5 shows the performance of the English and Dutch systems and their agreement. The rows indicate which feature groups were enabled (left column group). The centre column group contains the accuracy of the system, the percentage of predictions that matched to annotated value. The right column shows the *internal kappa*: this is the Cohen's kappa of the predictions of the English and Dutch systems. The human-annotated relations do not matter for this value, it signifies only how similar the two systems predicted entailment, corrected for chance.

len	**st**	**ss**	**English**	**Dutch**	**Kappa**
+			62.75%	60.75%	0.5806
	+		60.88%	59.88%	0.6169
		+	63.13%	62.75%	0.5683
+	+		62.63%	61.63%	0.6166
+		+	63.75%	63.00%	0.6429
	+	+	62.50%	63.35%	0.6347
+	+	+	64.63%	63.38%	0.6215

Table 5.5: Accuracy of the system with different feature groups enabled or disabled. Kappa represents the agreement between the two systems.

No particular feature groups seem to have a particularly strong effect on accuracy, and in general adding more features increases the accuracy. The kappa varies around 0.6, but it must be noted that the highest accuracy is preferred, and thus we look only at the kappa with all feature groups enabled. The highest accuracy means that the system will have the best internal representation of semantic information, which is what we are interested in. In other cases it may be that both systems have a method which manages to do well but not actually represent semantic information. This will be discussed in more depth in Section 6.1.

When all features are enabled, there are 283 incorrectly classified sentence pairs for English and 293 for Dutch, overlapping at 219 pairs.

Of the 138 sentence pairs that the systems disagreed on, the English system was right 74 times and the Dutch system was right 64 times.

5.2.2 Data set variables

Table 5.6 shows the distribution of the different variables in the data set for the sentence pairs that were predicted incorrectly for both systems, as discussed in Section 3. Looking at the incorrect prediction we can hopefully see where the system went wrong and why.

Text length

Lang.	Long	Short
English	43 (15%)	240 (85%)
Dutch	41 (14%)	252 (86%)

Sentence pair category

Lang.	IE	IR
English	94 (33%)	65 (23%)
Dutch	93 (32%)	74 (25%)

	QA	SUM
English	47 (17%)	77 (27%)
Dutch	52 (18%)	74 (25%)

Annotated entailment relation

Lang.	Yes	No
English	89 (31%)	194 (69%)
Dutch	90 (31%)	203 (69%)

Table 5.6: Distribution of data set variables across incorrectly predicted sentence pairs (283 total for English, 293 total for Dutch).

The text length (first table) seems heavily biased towards short texts. This does not carry any significance to the fact that sentence pairs were answered incorrectly though. The large variation can be explained purely by the original distribution of long and short texts in the data set. Roughly 17% of the data set has been flagged as long, and 22 is almost 17% of 138, the total amount of different answers. The other variables are equally distributed (25%/25%/25%/25% for the sentence pair category and 51,25%/48,75% for annotated entailment relation) in the data set. We conclude that these variables do not explain the difference between the two systems.

In the final table we do see a bias towards sentence pairs where no entailment was annotated. This does not seem to relate to the difference between the two systems as the bias is present in both results. A simple explanation for this bias is that the SVM overestimated the probability of there being an entailment relation; in case of doubt it tends to guess that there is entailment. Systems that classify by making a binary division in a feature space, such as an SVM, will almost always be biased toward one of the answers (unless perfect accuracy is achieved).

Table 5.7 shows the distribution of the same variables in the sentences where the systems disagreed (a subset of the complete data set). By looking at the sentence pairs that the systems predicted differently we can hopefully see the difference between the two systems, explaining the kappa value.

Text length

Long	Short
22 (16%)	116 (84%)

Sentence pair category

IE	IR
13 (9%)	51 (37%)

QA	SUM
41 (30%)	33 (24%)

Annotated entailment relation

Yes	No
75 (54%)	63 (46%)

Table 5.7: Distribution of data set variables across the sentence pairs that the systems disagreed on (138 total).

The text length shows a similar distribution as before, matching the distribution in the whole data set. The sentence pair category also shows a very similar distribution, but the sentence pairs that came from the IE task occur a lot less in this set. This does not mean that IE sentence pairs were done well, we saw in Table 5.6 that a normal amount of IE pairs were answered incorrectly. What this means is that for the IE pairs, the system responded very similarly.

This likely means that for IE sentence pairs, the translations are very similar to the original questions. If it is the case that IE sentence pairs suffer less from changes during translation that affect RTE systems, it may proof fruitful for new data sets to focus on constructing sentence pairs that way. However, because care was taken during creation of the data set to make all sentence pairs have the same format it seems more likely that this is just coincidence.

The last table shows that the correct answer seems to have little effect on the difference between the two systems.

5.2.3 Features

Because the SVM is largely a black box-machine it is very difficult to establish what causes the differences between the two systems. However, since the SVM in both systems is trained in the same way, the difference can likely be explained through the difference in feature values.

The features for the test sets of both languages were extracted, as well as the subset for all sentence pairs that the systems disagreed on. Student's t-tests were performed on all features separately using all sentence pairs, to see whether they differed significantly. For those that did the effect size was computed using:

$$\frac{\mu_d - \mu_e}{(\mu_d + \mu_e)/2} \tag{5.1}$$

where μ_d and μ_e are the means of the Dutch and English feature respectively. You can see that we take the difference between the two means and divide over the average mean to compensate for the fact that the features have completely different scales (a normal value for the amount of words in the texts is around 25, while that the amount of adjectives in the text is typically around 1). You can also see that a positive effect size means that Dutch had a higher mean while a negative effect size means that English had a higher mean, for that particular feature. The effect sizes are shown in Figure 5.1.

We see that especially the second half of the features differ significantly across languages. The Manhattan and Euclidean vector distance between tf*idf and ormf representations of the sentences prove the exception. Most of the length features are not even shown, because they are insignificant. Insignificance across the length features is to be expected, as word counts are one of the least changing properties across translations - at least for English and Dutch. The last 11 features use vector representations that use word occurrence across the entire data set and will thus quickly be more affected by smaller differences across translation. Whether it represents semantic information better will be discussed in the next section.

Overall we see that 13 out of 32 features are significant ($\alpha \leq 0.05$) for the sentence pairs that the systems disagreed on. Three additional features reach this level for all features but not for the disagreed pairs. These features have been marked with an asterisk (*) in Figure 5.1.

The direction of the effect does not seem to have a particular bias. Within feature groups some patterns do arise though.

For the five significant length (*len*) features it goes both ways, and for three of these five the difference for the sentence pairs where the systems disagreed on the effect is not even significant.

For the surface text similarity (*st*) features we see an interesting pattern: from the Jaccard similarity up to the Cosine similarity the effect sizes are negative, and for the three statistical vector correlation coefficients (Pearson, Spearman and Kendall's Tau) the effect is positive. Because all these features are similarity measures between vector representations of the sentences, this means that for the former group the English sentences differ more in tf*idf represenation, while for the latter the Dutch sentences differ more. The major difference between these groups is that the former uses word counts relative to all the words in the sentences, while the latter use word counts relative to all words in the entire data set (due to the tf*idf representation).

This means that within sentence pairs the English sentences differ more, but when their representation takes into account all the words in the data set, the Dutch sentences differ more. This means that on a local scale, English has more variation in its words (the text and hypothesis differ more from each other) but on a global scale the Dutch data set actually has more variation.

One reason for the Dutch data set having more variation in a global sense is the preprocessing phase of the systems performing differently, especially in the synonym normalisation step. In the

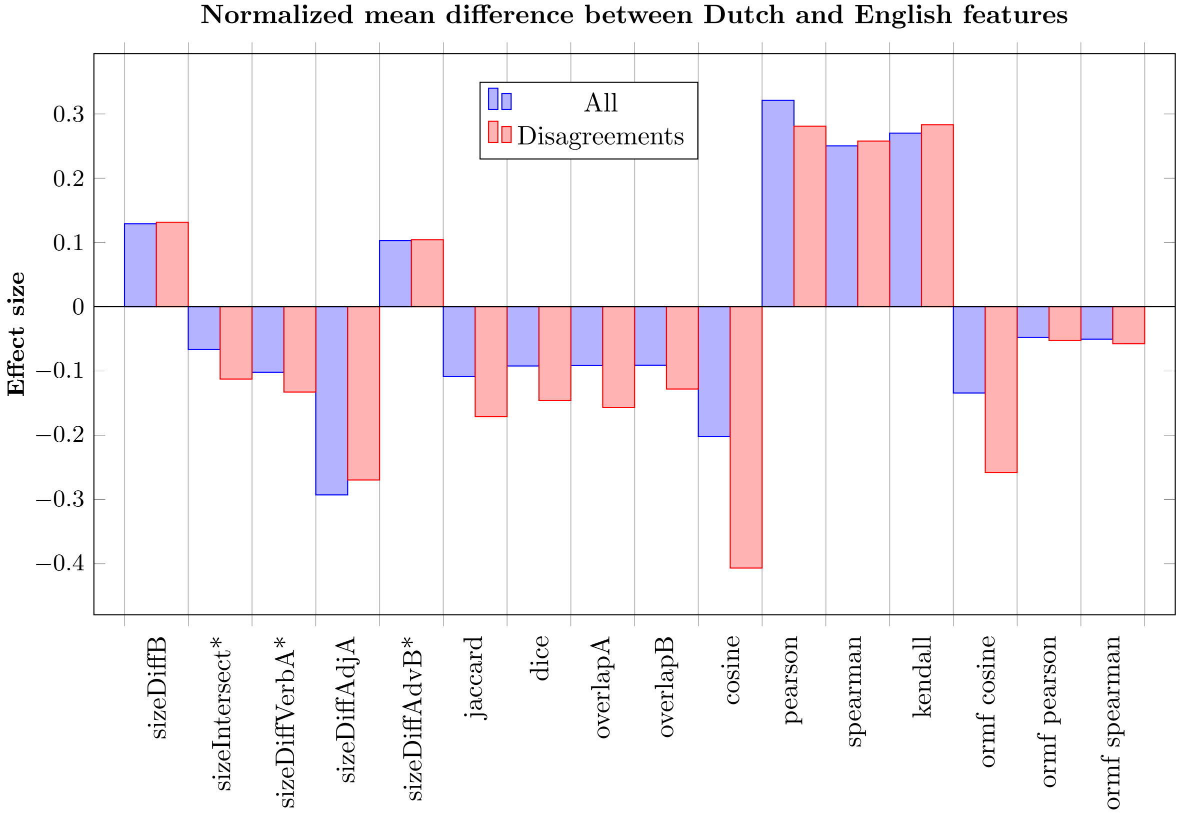

Figure 5.1: Effect sizes going from Dutch to English features. Only the features with a significant ($\alpha = 0.05$ difference are shown*. The blue bars represent the effect for all measured features while for the red bar only the features of the data points that the Dutch and English systems disagreed on were used. The effect size is computed according to Equation 5.1; a positive effect size means that the Dutch mean was higher and a negative effect size means that the English mean was higher.

*Features marked with an asterisk had a a significant difference for all sentence pairs but not for the set of sentence pairs where the systems disagreed.

English version, WordNet was used to find synonyms while for Dutch Open Dutch WordNet was used. Open Dutch WordNet is less developed than WordNet and it was noted during system development that the synsets were less complete. If less synonyms were found for Dutch, more variation in words was preserved while for English a lot of words were replaced by synonyms already existent somewhere else in the data set.

A reason for the Dutch data set having less variation in a local sense could be that the translation was done by a single person. One person has a limited vocabulary and tends to use certain words for synonymous or near-synonymous words. The English data set had various sources, texts written by all kinds of people and thus have a lot of ways of saying things. During translation the translator likely converged various ways of saying something into one, reducing variation within sentence pairs.

It is a bit unclear why this does not extend to the entire data set. The exact translation method is unknown, but it is assumable that the translator translated corresponding texts and hypotheses together and thus sentence pairs will be subjected to the same bias. It is also assumable that the translations were not made in one day and even within days it will have taken a significant amount of time. If we assume that the translator's bias changes over time it can explain why the reduced variation only applies to sentence pairs. It is however uncertain if this is the case. The translator's bias may also change dependent on the stimulus he receives, i.e. the English sentence and the order of sentence pairs.

The differences between the two halves of the surface text features are explained via this doubleedged knife: on one end you have convergence during translation decreasing the variation of words in the Dutch sentence pairs, relative to English, while on the other end you have a less effective synonym normalisation step that relatively increases variation within the Dutch data set.

For the semantic similarity (*ss*) features we see that all effect sizes are negative. The ORMF representation is reported by Zhao et al. (2014) to be good at modelling the meaning of words in the local context, so the negative effect size is in accordance with the theory mentioned above: English sentence pairs differ more on a local scale while Dutch sentence pairs differ more when represented relative to the entire data set.

This effect is surprising, and it may very well be a property of this specific data set. Recall that these features are only of the test set, meaning that these conclusions are based on 800 sentence pairs. It is recommended that for future research a larger data set is used as to prevent that we see extraneous effects as meaningful.

We see that the effect size for the sentence pairs where the systems disagreed is very often larger than for all sentence pairs, which supports the suspicion that the systems predicted differently because of differences in the features. When the features differ strongly, the systems start making different predictions, and some features seem to not make a difference while others do.

The cosine distance for both the tf*idf and ORMF representations of stentences stand out; the difference between these features appear much larger for the sentence pairs where the systems disagree and thus suggest that this is a major influence on the difference between the systems. According to Guo et al. (2014), the cosine similarity is one of the best similarity measures used in NLP, and we see here that it can strongly detect differences between sentence representations. Because the effect is so strong the cosine similarity will likely have a big influence on the difference between the systems, but as the SVM is a black-box machine we have no way of proving this.

This significance shows us that the same sentences are represented quite differently for the two languages. A small data sample and the fact that it is impossible to see into the decision making process of the SVM prevents us from attributing differences to specific features, but nonetheless we know that the systems for both languages will have trained differently and their decision making process will be different, likely incomparably so. The systems will have very different ways of predicting, yet both systems reach similar performance, in their own way. It is within the set of sentence pairs where the systems disagreed where we can really see that they have different decision making processes.

6 Discussion

6.1 Results analysis

Even though the full system acquired a kappa score of 0.6215 (Table 5.5), a higher agreement between the systems was found when the Surface Text Similarity feature group was disabled. In this case, the kappa score was 0.6429. Is this system better? Are we better off disabling that feature group altogether? This brings us to a much more important underlying question: *"Is semantic information actually being represented in the system?"*.

The system with only the length and semantic similarity groups enabled may have a higher kappa score, but it also has a lower accuracy in comparison. This accuracy represents how 'well' the system did; how much of its answers would be the same as human annotators. The kappa measures only how similar the systems answered, disregarding what would be correct answers. To illustrate, if both systems were simply set to always answers positively - or something similarly simple - the systems would have a kappa score of 1 while the accuracy would be around 50%. The kappa score was high, but no semantic information was represented internally. The systems never even looked at the sentences. So, a high kappa score is the aim, but the accuracy represents the reliability of a conclusion based on that score.

While the original ECNU system reached an accuracy of 83.6% for the RTE task in SemEval-2014 (Marelli et al., 2014), our accuracy around 64% is quite good. The data set used in SemEval-2014 was quite different so these numbers are not directly comparable, and it is incorrect to assume that the ECNU system can actually reach this 83%. Nonetheless, if the other 40 features defined in Zhao et al. (2014) are also implemented this will probably only improve and thus improve the meaning of the kappa score. Our result definitely surpasses chance and thus some semantic information must be represented.

It must be noted that so far all features look mostly at word counts, within the sentence pairs or within the entire data set. The use of POS tags, which are used in the second half of the length features, is one of the few examples of deeper information being used. Synonymity is also applied in the preprocessing phase. Overall, the features are quite superficial. The remaining 40 features also apply techniques like grammatical dependency, co-occurrence in various corpora, antonymity and WordNet distance. These features are considerably deeper, but due to time concerns and the difficulty of finding and using good Dutch alternatives for resources like the corpora and WordNet, these were not implemented.

The kappa score is already quite high, as it means not only that chance has little effect on the agreement, but also that the systems give the same answers on a large part of the sentence pairs, likely because the features actually represent something meaningful. As discussed in the previous section the systems likely have quite different methods of predicting the entailment relation, but still they reach similar results. For 138 sentence pairs one was correct while the other was not. If the manner in which semantic information is represented is improved (e.g. by adding more features) this set should grow smaller and smaller as both systems reach higher accuracy. If they both give the correct entailment relation, it does not matter if they arrived at the conclusion with different methods. Agreement should not represent similar methods, only similar results.

6.2 Further Research

We would like to openly invite anyone to continue this research by adding more features to the system, trying different machine learning algorithms or doing a similar thing for other languages. The code for this system is available,[9] together with the data set. Instructions for the remaining features can be found in Zhao et al. (2014), but of course other features can be added. Our experience has taught us that it is often easier to find an English resource than a Dutch one, so deviations from the original features may be made where a Dutch resource is the original and an English alternative has to be found rather than vice versa.

It may also prove useful to investigate other Machine Learning methods than SVM. As described in Section 2.3 the ECNU system used multiple classifiers with slightly different results. It would be interesting to see if all these classifiers perform similarly across translation, especially in relation to each other. In English and on the SemEval-2014 data set the SVM performed best, but it may be that when using a Dutch data set other classifiers suddenly become better.

It might also be useful to explore classifiers that give more insight in the classification process. The SVM proved to be a real black-box machine, making it very hard to identify what made the systems answer differently for certain sentence pairs. Using classifiers like kNN or decision trees more insight may be given in what features are important in both languages, and most importantly where they differ.

As we mentioned in Section 2 the ECNU system might have used some non-semantic properties of the SICK data set (used in SemEval-2014) to predict the entailment relation. We feel like without implementing all features it is not possible to say whether this causes the lower accuracy (in relation to their reported 83.6%). It also cannot explain why the systems do relatively well, because the exploitable properties are not in the data set used here. Nothing definitive can be said about this now, but this point should be kept in mind for later research as well.

As we have proven that data sets for RTE can be translated, at least between English and Dutch, all the ideas for further research mentioned above can

[9] `https://github.com/Superkebabbie/Translated-RTE`

hopefully be executed using larger data sets. 1.600 sentence pair is not a bad sample size, but some smaller effects were observed in Section 5.2 that we could not make a definitive conclusion about because of the sample size. Testing the system on a translated version of the SemEval-2014 data set would remove any concerns due to the different data sets and should guarantee us the accuracy of 83.6%. However, it must be noted that translating these sentence pairs is a lot of work and manual translation is not flawless (as seen in question 17 of the data set validation, Section 4.2).

6.3 Conclusion

The kappa score (0.6215) is quite high and suggests a positive result. However the accuracy, a value not corrected for chance, could be a fair bit higher and makes any conclusion based on this system somewhat doubtful, as we are unsure whether semantic information is properly represented. We would like to see the remaining 40 features added before any definitive conclusions. The preliminary conclusion is that the translated systems does perform as well on translated data as the original system does on English data. This suggests that semantic differences between Dutch and English are small and prove little problem for CLTE.

We have definitively proven that the translated data set is suitable for further research. It has shown that some translations are flawed, and more importantly that some sentence pairs were ambiguous even before translation, but these pairs do not occur frequently enough to form a serious problem for agreement tests[10].

This means that it is actually possible to translate existing data sets for English RTE studies, for use in RTE studies in other languages. This is very preferable over having to make data sets for each language independently. It also means that if data sets are made for other languages they can likely also be translated to English, and others language. It is reasonable to assume that if semantic information stays the same during translation from English to Dutch it also stays the same when translated from Dutch to English. Languages that differ more from English than Dutch does (i.e. non-Germanic languages) may be harder to translate while keeping semantic information the same, so further investigation in those languages is recommended.

Overall, this is a very positive result for such an exploratory study. We have preliminarily proven that RTE systems developed for English may be adapted for use in Dutch and probably other languages, and we have confirmed that one of the few - if not the only - data set available for RTE in Dutch is ready for use. We have also produced a system that is easy to continue this research with, and suggested a number of ideas for further research that it can be used for.

[10] The data set will be improved in the near future, in order to remove these flawed data points by the original translators.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics, 2012.

Johan Bos, Kilian Evang, Johannes Bjerva, Hessel Haagsma, Valerio Basile, and Noortje Venhuizen. The Parallel Meaning Bank: Annotation manual, version 0.3. `http://www.let.rug.nl/bos/pubs/pmbmanual2016.pdf`, 2016. Accessed: 06-07-2016.

Gosse Bouma, Gertjan Van Noord, and Robert Malouf. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37 (1):45–59, 2001. ISSN 0921-5034. URL `http://www.ingentaconnect.com/content/rodopi/lang/2001/00000037/00000001/art00004`.

Ido Dagan, Oren Glickman, and Magnini Bernardo. The PASCAL recognising textual entailment challenge. In J. Quionero-Candela, I. Dagan, B. Magnini, and F. d'Alch Buc, editors, *Machine Learning Challenges*, volume 3944, pages 177–190. Springer, 2006. URL `http://u.cs.biu.ac.il/~nlp/RTE1/Proceedings/dagan_et_al.pdf`.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Claudia Leacock. The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. *Language Resources and Evaluation*, 50(1):67–93, 2016. URL http://dx.doi.org/10.1007/s10579-015-9313-8.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 1–9, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1654536.1654538.

Weiwei Guo and Mona Diab. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 864–872. Association for Computational Linguistics, 2012.

Weiwei Guo, Wei Liu, and Mona T Diab. Fast tweet retrieval with compact binary codes. In *COLING*, pages 486–496. Citeseer, 2014.

Alice Lai and Julia Hockenmaier. Illinois-LH: A denotational and distributional approach to semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334, 2014.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

Bernardo Magnini. Recognizing textual entailment: Models and applications. *Computational Linguistics*, 41(1):157 – 159, 2015. ISSN 08912017.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *The 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, 2014.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. SemEval-2012 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 399–407. Association for Computational Linguistics, 2012.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, pages 300–308, Bucharest, Romania, 2016.

P. Vossen, I. Maks, R. Segers, H. van der Vliet, M-F. Moens, K. Hofmann, E. Tjong Kim Sang, and M. de Rijke. Cornetto: a combinatorial lexical semantic database for Dutch. In Jan Spyns, Peter; Odijk, editor, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme.*, Theory and Applications of Natural Language Processing, 2013, XVII, chapter 10. 2013. URL http://wordpress.let.vupr.nl/cornetto/.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78, 2014.

Jiang Zhao, Tiantian Zhu, and Man Lan. ECNU: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 271–277, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/S14-2044.