A Search Task Dataset for German Textual Entailment

Britta D. Zeller and Sebastian Padó
Department of Computational Linguistics, Heidelberg University, Germany
{zeller,pado}@cl.uni-heidelberg.de

Abstract

We present the first freely available large German dataset for Textual Entailment (TE). Our dataset builds on posts from German online forums concerned with computer problems and models the task of identifying relevant posts for user queries (i.e., descriptions of their computer problems) through TE. We use a sequence of crowdsourcing tasks to create realistic problem descriptions through summarisation and paraphrasing of forum posts. The dataset is represented in RTE-5 Search task style and consists of 172 positive and over 2800 negative pairs. We analyse the properties of the created dataset and evaluate its difficulty by applying two TE algorithms and comparing the results with results on the English RTE-5 Search task. The results show that our dataset is roughly comparable to the RTE-5 data in terms of both difficulty and balancing of positive and negative entailment pairs. Our approach to create task-specific TE datasets can be transferred to other domains and languages.

1 Introduction

Textual Entailment (TE) is a binary relation between two utterances, a *Text T* and a *Hypothesis H*, which holds if "a human reading T would infer that H is most likely true" (Dagan et al., 2005). Example 1 shows a positive entailment (T entails H_1) and a negative entailment (T does not entail H_2).

(1) **T:** Yoko Ono unveiled a bronze statue of her late husband, John Lennon, to complete the official renaming of England's Liverpool Airport as Liverpool John Lennon Airport.

H_1: Yoko Ono is John Lennon's widow.

H_2: John Lennon renamed Liverpool Airport.

The appeal of Textual Entailment is that it can arguably meet a substantial part of the semantic processing requirements of a range of language processing tasks such as Question Answering (Harabagiu and Hickl, 2006), Information Extraction (Romano et al., 2006), or Summarisation (Harabagiu et al., 2007). Consequently, there is now a research community that works on and improves Textual Entailment technology. In this spirit, the main TE forum, the yearly Recognising Textual Entailment (RTE) Challenge, has created a number of datasets that incorporate the properties of particular tasks, such as Semantic Search in RTE-5 (Bentivogli et al., 2009) or Novelty Detection in RTE-7 (Bentivogli et al., 2011).

At the same time, work on RTE on has focused almost exclusively on English. There is at most a handful of studies on Textual Entailment in other languages, notably German and Italian (Wang and Neumann, 2008; Negri et al., 2009; Bos et al., 2009) as well as a study on cross-lingual entailment (Mehdad et al., 2010).[1] Consequently, virtually no TE technology is available for non-English languages. What is more, it is not clear how well existing algorithms for English RTE carry over to other languages, which might show very different types of surface variation from English. The same limitation exists in terms of genre/register. Virtually all existing datasets have been created from "clean" corpora – that is, properly tokenised, grammatical text, notably Wikipedia. Again, the question arises how well TE

[1] There is also a translation of the RTE-3 dataset into German, but it is so far unpublished, although available from http://www.dfki.de/~neumann/resources.html

algorithms would do on noisier genres like transcribed speech or user-generated content. Arguably, it would benefit the community to have a larger variety of datasets at hand for such investigations.

This paper reports our creation and analysis of a German dataset for TE that is derived from social media data, as is produced every day on a large scale by of non-professional web users. This type of data respects linguistic norms such as spelling and grammar less than traditional textual entailment datasets (Agichtein et al., 2008), which present challenges to semantic processing.

We concentrate on a search task on a computer user forum that deals with computer problems: given a problem statement formulated by a user, identify all relevant forum threads that describe this problem. We created queries for a sample of forum threads by crowdsourcing. We asked annotators to summarise the threads and to paraphrase the summaries to achieve high syntactic and lexical variability. The resulting summaries can be understood as queries (problem statements) corresponding to the original posts. The search for relevant posts given a query can be phrased as a TE problem as follows: queries are hypotheses that are entailed by forum posts (texts) T iff the forum post is relevant for the query (Peñas et al., 2008).

Plan of the paper. Section 2 defines the task in more detail and describes the rationale behind our definition of the crowdsourcing tasks. Section 3 provides a detailed analysis of the queries that were produced by crowdsourcing. Section 4 assesses the difficulty of the dataset by modelling it with the RTE system EDITS (Kouylekov and Negri, 2010). Finally we relate our study to prior work and sum up.

2 Creating a German Social Media TE Dataset with Crowdsourcing

2.1 Rationale

As mentioned above, the promise of TE lies in its ability to model NLP tasks. One of the best-established of these tasks is search, which has been a part of the RTE challenges since RTE-5 (Bentivogli et al., 2009). In this setup, given a query statement and a set of documents, a document is relevant if it entails the query. That is, the documents serve as candidate texts T for a hypothesis H given by the query. We apply this setup to social media texts that discuss computer problems. Our use case is that a user has a particular problem with their machine and wants to retrieve the set of relevant documents from computer problem forums. In terms of the entailment-based search task, the Ts are given by a corpus of German computer forum threads. More specifically, we use the first post of each thread, since an analysis showed that the first post usually contains the problem description. What is missing, however, are plausible queries (i.e., Hs). We create these queries by asking laypersons to summarise posts through Amazon Mechanical Turk (AMT) (Snow et al., 2008). This involves three steps:

Summarisation. Given the first post of a forum thread (T), summarise the content in one sentence (H*).

Paraphrasing. Paraphrase H* into another sentence (H) by changing both syntax and lexical choice.

Validation. Given original post (T) and paraphrased summary (H), assess if H correctly summarises T.

Step 1 maps documents onto potential queries; these queries might however be still very close to the original verbalisation in the document. On the semantic level, we assume that summarisation can lose information, but not create new information; thus, summaries should be entailed by the original texts (Harabagiu et al., 2007). Step 2 allows that there is an amount of syntactic and lexical variance between T and H that is realistic for a search task. On the semantic level, we assume that paraphrasing preserves information; that is, input and output of this step should generally exhibit a high degree of semantic equivalence. Finally, Step 3 allows us to detect and remove bad queries produced by unmotivated or sloppy turkers. Thus, queries validated by Step 3 will be entailed by the original documents.

2.2 Crowdsourcing Details

We sampled 25 first posts of threads from a corpus of German computer self-help forums as Ts, each for which we generate several Hs. The posts were selected so that their length matches the distribution over lengths for all first posts in the corpus. All 25 posts have a length between 50 and 120 words.

	ps	is	ns
ps	168	211	47
is	0	132	87
ns	0	0	36

Table 1: Confusion matrix for pairs of AMT validation annotations

Task 1: Summarisation. In the first step, we asked AMT workers to write a concise summary of a forum post, summarising the central points of the original text in a declarative sentence. We also provide an example text with summary. Turkers could mark a text as unsummarisable, but had to indicate a reason.

The task was conducted by five turkers for each forum post, leading to $25 * 5 = 125$ potential summaries. Two posts were discarded as unsummarisable since they referred heavily to another forum post, which left us with 115 summaries. We paid 0.20 USD for each summary. (Total: 23 USD)

Task 2: Paraphrasing. In this task, workers had to reformulate the summaries produced in the first task. They were asked to replace words by appropriate synonyms and to change the sentence structure, while still maintain the meaning of the original sentence. The workers of Task 2 were not shown the original forum posts, only the summaries. Again, there was the possibility to leave the text unparaphrased, indicating a reason. Each sentence was paraphrased by two turkers, resulting in $115 * 2 = 230$ paraphrases.

We decided to discard four of the 230 paraphrases, including their two input sentences (summaries from Task 1). We found that these input sentences provide overly generic summaries of their posts to be usable. For example, a post which dealt with various strategies to solve pop-up problems in Firefox was summarised as “Mein Rechner öffnet selbstständig Webseiten [...].” (*“My computer opens web pages on its own [...].”*). We paid 0.10 USD for each of the 230 paraphrases. (Total: 23 USD)

Task 3: Validation. This task asked workers to judge whether the paraphrased summaries resulting from Task 3 are correct summaries of the problem described in T.[2] Possible answers were (a) perfect summary (“ps”); (b) incomplete summary that is missing central concept (“is”); (c) no (“ns”). We also asked turkers to verify that the paraphrased summaries were complete, grammatical, declarative sentences. Each T/H pair was assessed by 3 turkers who were paid 0.05 USD for each assessment. (Total: 35 USD)

Surprisingly, the most frequently chosen category was not “is” (41% of all assessments), but “ps” (43%). About 16% of the paraphrases summaries are judged as “ns”. To assess reliability, we computed a confusion matrix. In our three-annotation AMT setup where annotators are not necessarily constant across sentences, we decided to count the three pairwise annotations (*a1-a2, a2-a3, a1-a3*) for each sentence. Since the order of the annotators is random, we normalised to the order “ps” < “is” < “ns”. Table 1 shows the results. Satisfactorily, the diagonal, corresponding to matching judgements, shows the highest numbers. In total, 49% of the judgement pairs agree. The largest group of disagreements is “ps”/“is”; the number of “is”/“ns” cases is lower by a factor of two, and the number of “ns”/”ps” cases smaller by another factor of 2. We interpret these number as indication that the annotation task is fairly difficult, but that there is in particular a large number of clear correct cases. We build on this observation below.

2.3 Compilation of the Dataset

For each T/H pair, Task 3 provides us with three judgements on an ordinal scale with three categories: perfect summary (“ps”), incomplete summary (“is”), no summary (“ns”). The next question is how to select cases of true entailment and true non-entailment from this dataset.

Positive entailment pairs. As for entailment, we start by discarding all pairs that were tagged as “ns” by at least one rater. The situation is less clear for “is” cases: on one hand, hypotheses can drop information

[2] We used the term “summary” to describe the concept to our lay taggers which are unfamiliar with the term “entailment”.

Assessments	ps-ps-ps	ps-ps-is	ps-is-is	is-is-is	ns-ns-ns	ns-ns-is	ns-is-is
Entailment	Y	Y	Y	Y	N	N	N
Occurrence	38	45	50	20	7	11	21
Selected as (Non-)Entailment	37	41	42	7	7	2	1

Table 2: Association between AMT assessments and final entailment relations

present in the text while preserving entailment; on the other hand, the absence of important information in the summary can indicate difficulties with the original text or the summary. Thus, to keep precision high, we decided to manually check all "is"/"ps" T/H pairs. The left-hand part of Table 2 shows that in fact, the ratio of proper entailments trails off from almost 100% for "ps-ps-ps" to about one third for "is-is-is". In total, we obtained 127 positive entailment pairs in this manner.

During the extraction, we noted that one of the 23 forum posts did not yield reliable assessments for any of its generated hypotheses and discarded it.

Negative entailment pairs. Negative entailment pairs come from two sources. First, "ns" T/H pairs are cases where turkers missed the semantic core of the original text. These cases might be particularly informative non-entailment pairs because they are near the decision boundary. For example, one author asks whether a virus can settle down on the motherboard. The corresponding generated hypothesis turned the question into a fact, stating that "My BIOS has been infected by a virus.". Again, we checked all pairs with at least one "ns" judgement by hand. As the right-hand side of Table 2 shows, we find the same pattern as for positive pairs: perfect non-entailment for instances with perfect agreement on "ns", and lower non-entailment ratio for increasing "is" ratio. Rejected pairs are e.g. very generic and fuzzy summaries or refer only to a minor aspect of the problem described in the forum. Unfortunately, this strategy only results in 10 negative entailment T/H pairs. The second source of negative pairs are combinations of verified Hs with "other" Ts, that is, Ts from which they were not created. In fact, we can pair each of the 137 validated distinct Hs with all other Ts, resulting in $21 * 137 = 2877$ additional non-entailment T/H pairs.

However, since the domain of computer problems is relatively narrow, a few post topics are so close to each other that generated hypotheses are entailed by multiple texts. While this effect is usually ignored in machine learning (Bergsma et al., 2008), our goal is a clean dataset. Therefore, we manually checked all cross-pairs with similar topics (e.g. virus attacks) for positive entailment relations. Indeed, we found hypotheses which were general enough to match other texts. We removed 45 such pairs from the negative entailment pairs and added them to the set of positive pairs.

In total, we obtained 172 positive and 2842 negative entailment T/H pairs for 22 Ts and 137 distinct Hs. At a cost of 82 USD, this corresponds to an average of 50 cents for each explicitly generated positive pair, but just 3 cents for each T/H pair in the complete dataset. From the 226 AMT-generated pairs, we use 56% as positive pairs and 4% as negative pairs. We discard the remaining, inconsistently judged, 40%.

2.4 Discussion

The three tasks vary in their nature and difficulty. As mentioned above, we paid more for Task 1 than for Task 2, since it involved authoring a text. The amount of time needed for the tasks confirms this assumption: Task 1 took about 80 seconds per annotation, Task 2 only about 60 seconds. In terms of difficulty, Task 1 seems to be the easier one, though: We removed only a small number of post summaries from Task 1, but had to disregard a number of paraphrases from Task 2 (cf. Section 2.3). We believe that two factors contribute to this observation: (a), it is easier to summarise a complete text than to paraphrase a sentence out of context; (b), we deliberately asked workers in Task 2 to introduce as much variance as possible, which can lead to somewhat unnatural statements. Finally, the assessment Task 3 is the fastest one, requiring only about 30 seconds per annotation.

Post/Summary ID	Example (German/English)	Phenomenon
1/1	Rechner mit Virus infiziert. – *Computer infected with virus.*	Incomplete sentence
1/2	Mein Rechner ist von einem Virus befallen. – *My computer is infected by a virus.*	Personal point of view, short summary
1/3	Der Virtumonde-Virus lässt sich nicht entfernen. – *The Virtumonde virus cannot be removed.*	Pseudo-passive
25/1	Ich möchte, dass mein Board dauerhaft auf GB LAN schalten. – *I want that my board permanently to switch to GB LAN.*	Ungrammatical sentence
25/3	Wie lässt sich bei einer GB-Netzwerkkarte ein Fallback auf 100mbit verhindern? – *How can a fallback to 100mbit in a GB network adapter be prevented?*	Question
20/2	Heute ist schon 4 mal beim aufrufen des p5q deluxe-sammelthreads mein trendmicro virenscanner angeschlagen er meldet den Virus: TSPY_ONLINEG.FXG was kann ich dagegen machen? – *Today while calling the p5q deluxe collective threads my trendmicro virus scanner has given mouth already 4 times it reports the virus: TSPY_ONLINEG.FXG what can i do against this?*	Long summary, writing errors

Table 3: Linguistic phenomena in summarisation task

Our results show that both with regard to positive and negative entailment, three consistent judgements are sufficient for an almost perfect guarantee of the respective relation (cf. Table 2), but only a comparatively small sample of our data fall into these categories (around 15% for positive and 3% for negative entailment, respectively). Creators of a dataset therefore have the option of either making up for this loss by starting with more initial data, which leads to a higher overall cost, or to perform a subsequent expert-driven manual pass over the inconsistent candidates, as we did.

3 Analysis of Created Data

This Section illustrates the properties and problems of each step.

3.1 Task 1: Summarisation

Linguistic properties. Table 3 illustrates some of the phenomena appearing in the summarisation task, which seem to be largely specific to the particular genre (web forum texts) that we consider, while appearing less frequent in standard trainig data like newswire. Example 1/1 shows a typical "telegram style" summary which omits determiners and copula; Example 25/1 shows that not all summaries are even grammatical (underlined word). A comparison of examples 1/2 and 1/3 shows that the summaries either retain the personal point of view typically used by the original posts (using first-personal personal or possessive pronouns) or employ generic, impersonal formulations such as (pseudo-)passives. In one example, the AMT worker even cited the author of the original post using the introduction "Micha fragt, ob [. . .]" (*"Micha asks whether [. . .]"*). Similarly, 12 summaries use interrogative form (Example 25/3) like the original posts even though we explicitly asked the turkers to generate declarative sentences. Finally, example 20/2 illustrates typical writing errors, including the omission of punctuation and the defiance of German capitalisation rules. It is notable that turkers used this style, which is typically used for writing forum posts, even in the rather more formal AMT task environment. It occurs more frequently for original posts with the same style. Arguably, the turkers perceived this as the "correct" manner to summarise such posts, as our guidelines did not address this question.

Post/Summary/ Paraphrase ID	Example (German/English)	Phenomenon
2/1/1, 2/1/2	PC ⇒ Computer/Rechner *(computer)*	Abbreviation, loanword
10/2/1	CPU ⇒ Prozessor *(processor)*	Abbreviation
3/3/1	AntiVir *(specific anti-virus program)* ⇒ Anti-Viren-Programm – *anti-virus program*	Hypernym
9/5/2	starten – *to start* ⇒ booten – *to boot*	Synonym
5/4/2	wird Hilfe benötigt – *help is needed* ⇒ bedarf es Unterstützung – *support is required*	Support verb construction changes
8/3/2	Ich habe XP neu installiert – *I reinstalled XP* ⇒ Neuinstallation von XP – *Reinstallation of XP*	Nominalisation
13/5/2	starten – *to start* ⇒ gestartet werden – *to be started*	Active/passive switch
4/4/2	ich möchte [. . .] löschen – *I want to delete [. . .]* ⇒ [. . .] lässt sich nicht entfernen – *[. . .] cannot be removed* (literally: *does not let itself be removed*)	Change of perspective (pseudo-passive)
17/3/2	User frägt ob eine Schadsoftware sich auch in der Hardware einnisten kann. – *User asks if malware can also infect hardware.* ⇒ Kann die Hardware ebenfalls von Maleware befallen sein? – *Can hardware be affected by malware, too?*	Declarative/interrogative switch

Table 4: Linguistic phenomena in paraphrasing task

Content properties. Most summaries reproduced the original content correctly. The turkers apparently concentrated more on the content, i.e. writing a good summary, than formal task details, resulting, e.g. in interrogative formulations. This is not untypical for crowdsourcing tasks (Chen and Dolan, 2010).

Nonetheless, reproducing the context correctly was not trivial: some forum posts are rambling or vague and difficult to summarise. Summaries of such posts often either (a) do not cover the whole content or (b) are incorrect. Cases (a) lead to assessments of medium reliability in Task 3 (“H is an incomplete, but valid summary of T”). Cases (b) lead to negative entailment cases.

As intended, the results of Task 1 are significantly shorter than the original texts, with an average length of 11 words (min 3, max 39 words). Often, they use more general wording, e.g. “Der Prozessor läuft schnell heiß.” (*“The processor runs hot quickly”*) for a description containing a concrete temperature.

3.2 Task 2: Paraphrasing

Linguistic properties. In the paraphrasing task, workers were asked to change both syntax and word choice whenever possible. Although texts can contain many content words that are hard to paraphrase (e.g. basic level terms such as *table*), the problem is alleviated in the software domain where abbreviations and English loanwords that can be substituted easily are frequent (examples 2/1/1, 2/1/2, 10/2/1 in Table 4). The most frequent change was the replacement of verbs by synonyms and nouns by synonyms or hypernyms, as in examples 3/3/1 and 9/5/2. Some turkers modified both syntax and lexemes to vary support verb constructions (5/4/2).

While these phenomena are all “generic” paraphrasing devices that have been observed in previous studies on English and newswire text (Lin and Pantel, 2002; Bannard and Callison-Burch, 2005), we find two more classes of paraphrasing patterns that are specific to German and the social media domain, respectively. Prominent among German-specific changes are the large number of nominalisations (8/3/2) as well as active/passive switches (13/5/2). Next to the regular passive construction with the auxiliary *werden*, we often see “pseudo-passives” which use *lassen* combined with the reflexivised verb (4/4/2).

As for domain-specific patterns, we frequently observe the alternation of interrogative and declarative sentences (17/3/2) noted before which is caused by the tendency of the original posts to formulate problems as questions. Again, personalised and generic expressions alternate (4/4/2), which typically involves rephrasing first-person statements as third-person or impersonal ones – often though (pseudo-)passives.

The quality is generally higher in Task 2 than it is in Task 1. Although we asked the turkers to generate paraphrases by changing both syntax and lexis, they frequently modified just the syntax. However, this is not critical, since the summaries already exhibit varied word choice, so that there is enough variance between T and the corresponding true entailment Hs to avoid oversimplifying the TE task.

Content properties. Recall that no context was given in the paraphrasing task to avoid influencing the turkers with regard to vocabulary and syntax. In most cases, context was also not necessary. However, this also meant that some semantic errors occurred as a result of ambiguous formulations in summaries that were propagated into the paraphrase. For example, the author of one forum post explains that a BIOS update has failed and that he is afraid of restarting the computer. The corresponding summary "Fehlermeldung nach Bios-Update, Rechner trotzdem neustarten?" (*"Error message after Bios update, restart computer anyway?"*) is paraphrased with "Ich erhalte nach dem Update meines BIOS eine Fehlermeldung, soll ich den PC neu starten?" (*"I get an error message after the BIOS update, should I restart the PC?"*), which has rather the meaning of restarting the PC *in order to* overcome the problem. Consequently, the assessment in Task 3 was controversial (ps-is-ns, see Section 2.3) and lead to a rejection of the T/H pair. In the best case, such errors can also lead to clear rejections (ns-ns-ns).

A specific problem that we observed was the lack of domain knowledge by turkers. For example, the summary "Anschluss von einem zusätzlichem SATA-Gerät . . . " (*"Connection of an additional SATA device . . . "*) becomes "ich möchte Hardware von SATA . . . anschließen" (*"I want to connect hardware (made) by SATA . . . "*). This is an incorrect paraphrase: SATA is not a hardware manufacturer, but a type of interface. This problem extended to Task 3, where assessments were controversial (ps-is-ns).

Finally, some turkers, contrary to instructions, produced summaries of the summaries. These texts became very short and were often marked as "is" (valid but incomplete) in Task 3. We observed that it was mostly turkers who already participated in Task 1 who acted in this manner. We feel that there is a tension regarding re-employing workers who participated in previous tasks: quality may profit from their previous training, but suffer from their bias to approach the second task with the same mindset as the first one.

3.3 Task 3: Validation

The output of the validation task allows us to correlate the quality ratings of T/H pairs to their linguistic properties. We observe a broad overlap between assessments of the type "is" and hypotheses which are very short or whose content is very general, e.g. due to the usage of hypernyms. Accordingly, T/H pairs which are marked consistently as "ps" concern either hypotheses which are relatively comprehensive, or texts which describe rather simple situations. At the opposite end of the scale, T/H pairs with three "ns" assessments arise from to propagated errors. T/H pairs marked with all three categories, ps-is-ns, make up only about 3%. These cases frequently refer to posts with complex queries such as users describing a sequence of problems. Such posts are hard to summarise and to evaluate, but are also unlikely search queries. The average length of the Hs selected through Task 3 is 11.4 words (min 5, max 22).

In sum, we interpret the three-stage crowdsourcing task as a success: The first two tasks generate a broad variation of potentially true T/H pairs, while the third task enables a filtering of dubious pairs. Although the linguistic quality of the obtained hypotheses shows clear imperfections, the quality of the original texts is equally low: the resulting T/H pairs reflect particularities of the social media domain. Example 2 shows (part of) a T/H pair; note the ungrammaticality in both T and H.

(2) **T:** [...] Ich habe heute alles zusammengebaut, aber aheb folgende probleme... 1.Der PC brauch ca 5-10min zum booten. 2.Nach dem Starten hängt der pc sich ständig auf. [...] 4.beim booten wird "Pri Master Hard Disk : S.M.A.R.T. Status BAD, Backup and Replace Press F1 to Resume." wenn ich denn F1 drücke fährt der pc weiter hoch. MFG

	Accuracy	P	R	F_1
		for positive entailment		
Word overlap	.93	.38	.38	.38
EDITS (edit distance)	.95	.63	.34	.44

Table 5: Test set results on social media dataset for two simple Textual Entailment algorithms

[...] I have assembled everything today, but haev the following problems: 1.The PC take ca 5-10min to boot. 2.After starting the pc locks up constantly. [...] 4. while booting is "Pri Master Hard Disk : S.M.A.R.T. Status BAD, Backup and Replace Press F1 to Resume." than when I press F1 the pc continues booting. RSVP

H: Meinen Computer benötig für das Hochfahren sehr lange und zeigt mir dann eine Meldung für einen Fehler an.
Mine computer need a long time for booting and then shows me a message for an error.

4 Modelling the Dataset with Textual Entailment Systems

In order to evaluate the difficulty of the dataset that we have created, we performed experiments with two different TE engines. We split our dataset into a development and a test set. Both sets are identical in terms of size (1507 T/H pairs) and amount of positive and negative pairs (86 and 1421 pairs, respectively).

The first system is EDITS (Negri et al., 2009), version 3.0.[3] EDITS uses string edit distance as a proxy of semantic similarity between T and H and classifies pairs as entailing if their normalised edit distance is below a threshold θ which can be optimised on a development set. While additional entailment knowledge can be included, no such knowledge is currently available for German and we use the default weights. The second system is a simple word overlap strategy which approximates semantic similarity through the fraction of H words that also occur in T (Monz and de Rijke, 2001). Again, pairs are classified as entailing if this fraction is larger than a threshold θ.

We preprocessed the data by lemmatising it with TreeTagger (Schmid, 1994) and removing stop words, employing a German stop word list which includes keywords from the social media domain.[4] The thresholds θ for both systems were set by optimising the F_1 score for positive entailment on the train set.

Table 5 shows the results for the word overlap model and EDITS. The very high accuracy values merely reflect the predominance of the negative entailment class; we therefore concentrate on the F-score statistics for positive entailment. We find that edit distance outperforms word overlap with F_1 scores of .44 and .38, respectively. Since the main difference between the two approaches is that edit distance is sensitive to word order, order information appears to be indeed informative: reordering between T and H do not incur costs in the word overlap model, but they do in the edit distance model. Example 3 shows a T/H pair with high word overlap, but negative entailment. It is correctly classified by EDITS, but misclassified by the word overlap model.

(3) **T:** Hallo PC-Freunde, ich habe letzte Woche XP neu installiert. Heute ist mir aufgefallen das die CPU-Auslastung immer zwischen 60% und 80% liegt obwohl im Taskmanager der Lerlaufprozess mit 90-99% angezeigt wird. Kann es vieleicht sein das im Taskmanager nicht alle Programme erfasst werden(währe mir neu) oder könnte vieleicht ein Virus, Trojaner sein der diese ununterbrochen hohe Auslastung bewirkt? Vobei mein Antivirusprogramm (Awast) keinen Virus oder ähnliches erkennt. [. . .]
[. . .] Today I realised that the CPU load is always between 60% and 80% although the idle task is always displayed with 90-99% in the task manager. Is it mabe possible thet not all

[3] Downloadable from `http://sourceforge.net/projects/edits/files/`

[4] `http://solariz.de/649/deutsche-stopwords.htm`

programs are captured in the task manager(whould be new to me) or could mabe be a virus, trojan horse which causes this steadily high load? Hovever my anti virus program (Awast) does not recognise a virus or the like. [. . .]

H: Die Prozessorauslastung ist bei 100% und Antivirenprogramme funktionieren nicht.
The processor load is at 100% and anti virus programs do not work.

Example 4 shows the opposite case, namely a positive T/H entailment pair that hardly shares any vocabulary since many T details are omitted in H. Both systems are unable to correctly label this instance.

(4) **T:** Es gibt bei m ir zwei Probleme bei der Ausführung des Tools unter Vista. 1) Vista blockiert die Ausführung mit dem Kommentar " ...Sie verfügen eventuell nicht über ausreichende Berechtigungen... " und 2) F-Secure gibt eine Malware-Warnung aus " W32/ Suspicious_U.gen " Virus. Ist die Viruswarnung nur ein Fehlalarm?
I h ave two problems with the execution of the tool under Vista. 1) Vista blocks the execution with the comment " ...You might not have sufficient authorisation... " and 2) F-Secure gives a malware warning " W32/ Suspicious_U.gen " Virus. Is the virus warning just a false alarm?

H: Wegen fehlenden Systemrechten des Anwenders in Windows kann die Datei nicht gestartet werden. – *The file cannot be started due to missing system rights by the user in Windows.*

The most direct point of comparison for our dataset is the RTE-5 search pilot (Bentivogli et al., 2009). The two main differences are language (English vs. German) and genre (newswire vs. social media). We found our dataset to be slightly easier to model. Part of the reason is the somewhat more balanced positive/negative distribution in our dataset: a random baseline achieves an F-Score of 8.4% on RTE-5 and 10.4% on our data. However, the improvement of informed models is also somewhat higher: EDITS without additional knowledge resources achieves 32.6% F-Score on RTE-5 (+24% over the baseline) (Bentivogli et al., 2009) and 44% F-Score on our dataset (+34% over the baseline). We believe that this is due to the greater coherence of our dataset: it deals with just one topic, while the RTE-5 dataset covers ten topics. We also observe that the Hs in RTE-5 are shorter than ours (avg. length 8.75 words vs. 11.4) which presumably leads to worse sparsity problems. Nevertheless, the results on the two datasets for baselines and simple methods are still remarkably similar.

5 Related work

In the Textual Entailment community, particularly in the studies who create datasets and resources, there is a strong focus on the English language (Androutsopoulos and Malakasiotis, 2010). All RTE datasets, the most widely used experimental materials, are in English. A few datasets have been created for other languages. To our knowledge, only an Italian one (Bos et al., 2009) and a Spanish one are freely available (Peñas et al., 2006). Datasets for other languages have been created in the context of the CLEF QA Answer Validation and Machine Reading tasks, but do not appear to be available to the general community.

We have employed crowdsourcing, a technique whose practice has expanded greatly over the last years (Snow et al., 2008). It has rarely been used for Textual Entailment, though, since high-quality crowdsourcing relies on the ability to formulate the task in layman's terms, which is challenging for entailment. We avoided this problem by asking turkers to provide summaries and paraphrases in two separate steps. Wang and Callison-Burch (2010) also use crowdsourcing to collect hypotheses for TE. In contrast to us, they do not ask turkers for full summaries and paraphrases, but have them extract facts from texts and create counter-facts from facts by inserting negations, using antonyms, or changing adverbs.

Finally, Bernhard and Gurevych (2008) present a study on data that is similar to ours. Their goal is the automatic collection of paraphrases for English questions on social Q&A sites. Employing similar methods to us (e.g., word overlap and edit distance), they achieve very good results. Their task is simpler in that in concentrates on paraphrase relations among statements rather than summarisation relations between texts and statements.

6 Conclusions

This paper makes two contributions. The first one is a freely available dataset[5] for Textual Entailment tasks which covers (a) a new language, namely German; and (b), a new genre, namely web forum text. The dataset models a search task on web forums, with short queries as hypotheses and forum posts as text candidates. Being constructed from real social media data, our data is more noisy than existing RTE datasets and shows novels linguistic paraphrasing phenomena such as switches between interrogative and declarative sentences. We consider our dataset to be a test bed for TE algorithms that have to deal with spontaneous and sloppy language, e.g. for other social media areas or on transcribed spoken language.

Our second contribution is a crowdsourcing-based procedure to create the dataset which can be applied to other languages and data sources in order to create comparable datasets quickly and at modest expense. The three-step setup that we introduce consists of a summarisation step, a paraphrasing step, and a validation step. This setup guarantees syntactic and lexical variation and makes it possible to detect and remove the sizable portion of the data that consists of queries that are either invalid or hard to judge. The number of summaries and paraphrases can be chosen according to the requirements of the dataset; as for validation, we found that three judgments were sufficient for a final categorisation. An alternative to our rather artificial way to collect data is presented in (Baldwin et al., 2010), employing web forum structure.

We have presented an experiment with two basic TE algorithms which establishes that the difficulty of the dataset is roughly comparable with the RTE-5 Search task testset. However, both algorithms were essentially knowledge-free, and we will conduct experiments with more informed algorithms. We expect the inclusion of lexical entailment knowledge (such as hyponymy relations) to provide a clear benefit. However, the top systems on the RTE-5 Search-Task, where the best result was 46% F-Score (+13% F-Score over edit distance) crucially employed lexico-syntactic paraphrase knowledge à la DIRT (Lin and Pantel, 2002). It remains to be seen how such syntax-based TE algorithms do on our dataset, where we expect parsing results to be substantially more noisy than for traditional RTE datasets.

Acknowledgments. This work was supported by the EC project EXCITEMENT (FP7 ICT-287923).

References

Agichtein, E., C. Castillo, D. Donato, A. Gionis, and G. Mishne (2008). Finding high-quality content in social media. In *Proceedings of WSDM*, Stanford, CA, pp. 183–194.

Androutsopoulos, I. and P. Malakasiotis (2010). A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research 38*, 135–187.

Baldwin, T., D. Martinez, R. B. Penman, S. N. Kim, M. Lui, L. Wang, and A. MacKinlay (2010). Intelligent linux information access by data mining: the ILIAD project. In *Proceedings of the NAACL Workshop on Computational Linguistics in a World of Social Media*, Los Angeles, CA, pp. 15–16.

Bannard, C. and C. Callison-Burch (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*, Ann Arbor, MI, pp. 597–604.

Bentivogli, L., P. Clark, I. Dagan, H. Trang Dang, and D. Giampiccolo (2011). The seventh PASCAL recognising textual entailment challenge. In *Proceedings of TAC*, Gaithersburg, MD.

Bentivogli, L., I. Dagan, H. T. Dang, D. Giampiccolo, M. L. Leggio, and B. Magnini (2009). Considering discourse references in textual entailment annotation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pisa, Italy.

Bentivogli, L., B. Magnini, I. Dagan, H. Trang Dang, and D. Giampiccolo (2009). The fifth PASCAL recognising textual entailment challenge. In *Proceedings of TAC*, Gaithersburg, MD.

[5] Can be downloaded from `http://www.excitement-project.eu/`.

Bergsma, S., D. Lin, and R. Goebel (2008). Discriminative Learning of Selectional Preference from Unlabeled Text. In *Proceedings of EMNLP*, Honolulu, Hawaii, pp. 59–68.

Bernhard, D. and I. Gurevych (2008). Answering learners' questions by retrieving question paraphrases from social Q&A sites. In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, Ohio, pp. 44–52.

Bos, J., M. Pennacchiotti, and F. M. Zanzotto (2009). Textual entailment at EVALITA 2009. In *Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia.

Chen, D. L. and W. B. Dolan (2010). Building a persistent workforce on Mechanical Turk for multilingual data collection. In *Proceedings of the AAAI Human Computation Workshop*, San Francisco, CA.

Dagan, I., O. Glickman, and B. Magnini (2005). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, UK.

Harabagiu, S. and A. Hickl (2006). Methods for using textual entailment in open-domain question answering. In *Proceedings of COLING/ACL*, Sydney, Australia, pp. 905–912.

Harabagiu, S., A. Hickl, and F. Lacatusu (2007). Satisfying information needs with multi-document summaries. *Information Processing and Management 43*(6), 1619–1642.

Kouylekov, M. and M. Negri (2010). An open-source package for recognizing textual entailment. In *Proceedings of the ACL 2010 System Demonstrations*, Uppsala, Sweden, pp. 42–47.

Lin, D. and P. Pantel (2002). Discovery of inference rules for question answering. *Journal of Natural Language Engineering 7*(4), 343–360.

Mehdad, Y., M. Negri, and M. Federico (2010). Towards cross-lingual textual entailment. In *Proceedings of HLT/NAACL*, Los Angeles, CA, pp. 321–324.

Monz, C. and M. de Rijke (2001). Light-weight entailment checking for computational semantics. In *Proceedings of ICoS*, Siena, Italy, pp. 59–72.

Negri, M., M. Kouylekov, B. Magnini, Y. Mehdad, and E. Cabrio (2009). Towards Extensible Textual Entailment Engines: the EDITS Package. In *Proceeding of IAAI*, Reggio Emilia, Italy.

Peñas, A., Á. Rodrigo, V. Sama, and F. Verdejo (2008). Testing the reasoning for question answering validation. *Journal of Logic and Computation 18*, 459–474.

Peñas, A., Á. Rodrigo, and F. Verdejo (2006). SPARTE: a test suite for recognising textual entailment in spanish. In A. Gelbukh (Ed.), *Proceedings of CICLing*, Lecture Notes in Computer Science.

Romano, L., M. Kouylekov, I. Szpektor, I. Dagan, and A. Lavelli (2006). Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of EACL*, Trento, Italy, pp. 401–408.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of ICNLP*, Manchester, UK.

Snow, R., B. O'Connor, D. Jurafsky, and A. Ng (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, Honolulu, HI, pp. 254–263.

Wang, R. and C. Callison-Burch (2010). Cheap facts and counter-facts. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk*, pp. 163–167.

Wang, R. and G. Neumann (2008). Information synthesis for answer validation. In *Proceedings of CLEF 2008*, Aarhus, Denmark.