Acquiring entailment pairs across languages and domains: A data analysis

Manaal Faruqui
Dept. of Computer Science and Engineering
Indian Institute of Technology
Kharagpur, India
manaal.iitkgp@gmail.com

Sebastian Padó
Seminar für Computerlinguistik
Universität Heidelberg
Heidelberg, Germany
pado@cl.uni-heidelberg.de

Abstract

Entailment pairs are sentence pairs of a premise and a hypothesis, where the premise textually entails the hypothesis. Such sentence pairs are important for the development of Textual Entailment systems. In this paper, we take a closer look at a prominent strategy for their automatic acquisition from newspaper corpora, pairing first sentences of articles with their titles. We propose a simple logistic regression model that incorporates and extends this heuristic and investigate its robustness across three languages and three domains. We manage to identify two predictors which predict entailment pairs with a fairly high accuracy across all languages. However, we find that robustness across domains within a language is more difficult to achieve.

1 Introduction

Semantic processing has become a major focus of attention in NLP. However, different applications such as Question Answering, Information Extraction and Machine Translation often adopt very different, task-specific semantic processing strategies. Textual entailment (TE) was introduced by Dagan et al. (2006) as a "meta-task" that can subsume a large part of the semantic processing requirements of such applications by providing a generic concept of inference that corresponds to "common sense" reasoning patterns. Textual Entailment is defined as a relation between two natural language utterances (a Premise P and a Hypothesis H) that holds if "a human reading P would infer that H is most likely true". See, e.g., the ACL "challenge paper" by Sammons et al. (2010) for further details.

The successive TE workshops that have taken place yearly since 2005 have produced annotation for English which amount to a total of several thousand entailing Premise-Hypothesis sentence pairs, which we will call *entailment pairs*:

(1) **P:** Swedish bond yields end 21 basis points higher.
H: Swedish bond yields rose further.

From the machine learning perspective assumed by many approaches to TE, this is a very small number of examples, given the complex nature of entailment. Given the problems of manual annotation, therefore, Burger and Ferro (2005) proposed to take advantage of the structural properties of a particular type of discourse – namely newspaper articles – to automatically harvest entailment pairs. They proposed to pair the title of each article with its first sentence, interpreting the first sentence as Premise and the title as Hypothesis. Their results were mixed, with an average of 50% actual entailment pairs among all pairs constructed in this manner. SVMs which identified "entailment-friendly" documents based on their bags of words lead to an accuracy of 77%. Building on the same general idea, Hickl et al. (2006) applied a simple unsupervised filter which removes all entailment pair candidates that "did not share an entity (or an NP)". They report an accuracy of 91.8% on a manually evaluated sample – considerably better Burger and Ferro. The article however does not mention the size of the original corpus, and whether "entity" is to

be understood as named entity, so it is difficult to assess what its recall is, and whether it presupposes a high-quality NER system.

In this paper, we model the task using a logistic regression model that allows us to synchronously analyse the data and predict entailment pairs, and focus on the question of how well these results generalize across domains and languages, for many of which no entailment pairs are available at all. We make three main contributions: (a), we define an annotation scheme based on semantic and discourse phenomena that can break entailment and annotate two datasets with it; (b), we idenfiy two robust properties of sentence pairs that correlate strongly with entailment and which are robust enough to support high-precision entailment pair extraction; (c), we find that cross-domain differences are actually larger than cross-lingual differences, even for languages as different as German and Hindi.

Plan of the paper. Section 2 defines our annotation scheme. In Section 3, we sketch the logistic regression framework we use for analysis, and motivate our choice of predictors. Sections 4 and 5 present the two experiments on language and domain comparisons, respectively. We conclude in Section 6.

2 A fine-grained annotation scheme for entailment pairs

The motivation of our annotation scheme is to better understand why entailment breaks down between titles and first sentences of newswire articles. We subdivide the general *no* entailment category of earlier studies according to an inventory of reasons for non-entailment that we collected from an informal inspection of some dozen articles from an English-language newspaper. Additionally, we separate out sentences that are ill-formed in the sense of not forming one proposition.

2.1 Subtypes of non-entailment

No-par (Partial entailment). The Premise entails the Hypothesis almost, but not completely, in one of two ways: (a), The Hypothesis is a conjunction and the Premise entails just one conjunct; or (b), Premise and Hypothesis share the main event, but the Premise is missing an argument or adjunct that forms part of the Hypothesis. Presumably, in our setting, such information is provided by the other sentences in the article than the first one. In Ex. (1), if P and H were switched, this would be the case for the size of the rise.

No-pre (Presupposition): The Premise uses a construction which can only be understood with information from the Hypothesis, typically a definite description or an adjunct. This category arises because the title stands before the first sentence and is available as context. In the following example, the Premise NP "des Verbandes" can only be resolved through the mention of "VDA" (the German car manufacturer's association) in the Hypothesis.

(2) **P:** Herzog wird in dem vierköpfigen Führungsgremium des Verbands für die Teile-
Herzog will in the four-head management board of the association for the parts
und Zubehörindustrie zuständig sein.
and accessory business resposible be.

H: Martin Herzog wird VDA-Geschäftsführer.
Martin Herzog becomes VDA manager.

No-con (Contradiction): Direct contradiction of Premise and Hypothesis.

(3) **P:** Wie die innere Uhr [...] funktioniert, ist noch weitgehend unbekannt.
How the biological clock [...] works, is still mostly unknown.

H: Licht stellt die innere Uhr.
Light regulates the biological clock.

No-emb (Embedding): The Premise uses an embedding that breaks entailment (e.g., modal adverbials or non-factural embedding verb). In the following pair, the proposition in the Hypothesis is embedded under "expect".

(4) **P:** An Arkansas gambling amendment [...] is expected to be submitted to the state Supreme Court Monday for a rehearing, a court official said.
H: Arkansas gaming petition goes before court again Monday

No-oth (Other): All other negative examples where Premise and Hypothesis are well-formed, and which could not be assigned to a more specific category, are included under this tag. In this sense, "Other" is a catch-all category. Often, Premise and Hypothesis, taken in isolation, are simply unrelated:

(5) **P:** Victor the Parrot kept shrieking "Voda, Voda" – "Water, Water".
H: Thirsty jaguar procures water for Bulgarian zoo.

2.2 Ill-formed sentence pairs

Err (Error): These cases arise due to errors in sentence boundary detection: Premise or Hypothesis may be cut off in the middle of the sentence.

Ill (Ill-formed): Often, the titles are not single grammatical sentences and can therefore not be interpreted sensibly as the Hypothesis of an entailment pair. They can be incomplete proposition such as NPs or PPs ("Beautiful house situated in woods"), or, frequently, combinations of multiple sentences ("RESEARCH ALERT - Mexico upped, Chile cut.").

3 Modeling entailment with logistic regression

We will model the entailment annotation labels on candidate sentence pairs using a logistic regression model. From a machine learning point of view, logistic regression models can be seen as a rather simple statistical classifier which can be used to acquire new entailment pairs. From a linguistic point of view, they can be used to explain the phenomena in the data, see e.g., Bresnan et al. (2007).

Formally, logistic regression models assume that datapoints consist of a set of predictors x and a binary response variable y. They have the form

$$p(y = 1) = \frac{1}{1 + e^{-z}} \text{ with } z = \sum_i \beta_i x_i \tag{1}$$

where p is the probability of a datapoint x, β_i is the coefficient assigned to the linguistically motivated factor x_i. Model estimation sets the parameters β so that the likelihood of the observed data is maximized.

From the linguistics perspective, we are most interested in analysing the importance of the different predictors: for each predictor x_i, the comparison of the estimated value of its coefficient β_i can be compared to its estimated standard error, and it is possible to test the hypothesis that $\beta_i = 0$, i.e., the predictor does not significantly contribute to the model. Furthermore, the absolute value of β_i can be interpreted as the *log odds* – that is, as the change in the probability of the response variable being positive depending on x_i being positive.

$$e^{\beta_i} = \frac{P(y = 1|x = 1, \dots)/P(y = 0|x = 1, \dots)}{P(y = 1|x = 0, \dots)/P(y = 0|x = 0, \dots)} \tag{2}$$

The fact that z is just a linear combination of predictor weights encodes the assumption that the log odds combine linearly among factors.

From the natural language processing perspective, we would like to create predictions for new observations. Note, however, that simply assessing the significance of predictors on some dataset, as

provided by the logistic regression model, corresponds to an evaluation of the model on the training set, which is prone to the problem of overfitting. We will therefore in our experiments always apply the models acquired from one dataset on another to see how well they generalize.

3.1 Choice of Predictors

Next, we need a set of plausible predictors that we can plug into the logistic regression framework. These predictors should ideally be language-independent. We analyse the categories of our annotation, as an inventory of phenomena that break entailment, to motivate a small set of robust predictors.

Following early work on textual entailment, we use word overlap as a strong indicator of entailment (Monz and de Rijke, 2001). Our **weighted overlap** predictor uses the well-known tf/idf weighting scheme to compute the overlap between P and H (Manning et al., 2008):

$$weightedOverlap(T, H, D) = \frac{\sum_{w \in T \cap H} \text{tf-idf}(w, D)}{\sum_{w \in H} \text{tf-idf}(w, D)} \quad (3)$$

where we treat each article as a separate document and the whole corpus as document collection D. We expect that No-oth pairs have generally the lowest weighted overlap, followed by No-par pairs, while Yes pairs have the highest weighted overlap. We also use a categorical version of this observation in the form of our **strict noun match** predictor. This predictor is similar in spirit to the proposal by Hickl et al. (2006) mentioned in Section 1. The boolean strict noun match predictor is true if all Hypothesis nouns are present in the Premise, and is therefore a predictor that is geared at precision rather than recall. A third predictor that was motivated by the No-par and No-oth categories was the number of words in the article: No-oth sentence pairs often come from long articles, where the first sentence provides merely an introduction. For this predictor, **log num words**, we count the total number of words in the article and logarithmize it.[1] The remaining subcategories of No were more difficult to model. No-pre pairs should be identifiable by testing whether the Premise contains a definite description that cannot be accommodated, a difficult problem that seems to require world knowledge. Similarly, the recognition of contradictions, as is required to find No-con pairs, is very difficult in itself (de Marneffe et al., 2008). Finally, No-emb requires the detection of a counterfactual context in the Premise. Since we do not currently see robust, language-independent ways of modelling these phenomena, we do not include specific predictors to address them.

The situation is similar with regard to the Err category. While it might be possible to detect incomplete sentences with the help of a parser, this again involves substantial knowledge about the language. The Ill category, however, appears easier to target: at least cases of Hypotheses consisting of multiple phrases case be detected easily by checking for sentence end markers in the middle of the Hypothesis (full stop, colon, dash). We call this predictor **punctuation**.

4 Experiment 1: Analysis by Language

4.1 Data sources and preparation

This experiment performs a cross-lingual comparison of three newswire corpora. We use English, German, and Hindi. All three belong to the Indo-European language family, but English and German are more closely related.

For English and German, we used the Reuters RCV2 Multilingual Corpus[2]. RCV2 contains over 487,000 news stories in 13 different languages. Almost all news stories cover the business and politics domains. The corpus marks the title of each article; we used the sentence splitter provided by Treetagger (Schmid, 1995) to extract the first sentences. Our Hindi corpus is extracted from the text collection of South Asian languages prepared by the EMILLE project (Xiao et al., 2004)[3]. We use the Hindi

[1] This makes the coefficiently easier to interpret. The predictive difference is minimal.

[2] http://trec.nist.gov/data/reuters/reuters.html

[3] http://www.elda.org/catalogue/en/text/W0037.html

No. of sentence pairs	English	German	Hindi
Original	473,874 (100%)	112,259 (100%)	20,209 (100%)
Filtered	264.711 (55.8%)	50.039 (44.5%)	10.475 (51.8%)

Table 1: Pair extraction statistics

Corpus	err	ill	no-con	no-emb	no-oth	no-par	no-pre	yes
English Reuters	3.5	2.9	0	0.2	3.7	7.4	0	82.3
German Reuters	2.1	11.0	0.4	0.2	4.3	2.1	0.2	79.7
Hindi Emille	1.1	2.5	0	0.3	14.7	5.7	0	75.7

Table 2: Exp.1: Distribution of annotation categories (in percent)

monolingual data, which was crawled from Webdunia,[4] an Indian daily online newspaper. The articles are predominantly political, with a focus on Indo-Pakistani and Indo-US affairs. We identify sentence boudaries with the Hindi sentence marker ('।'), which is used exclusively for this purpose.

We preprocessed the data by extracting the title and the first sentence, treating the first sentence as Premise and the title as Hypothesis. We applied a filter to remove pairs where the chance of entailment was impossible or very small. Specifically, our filter keeps only sentence pairs that (a) share at least one noun and where (b) both sentences include at least one verb and are not questions. Table 1 shows the corpus sizes before and after filtering. Note that the percentage of selected sentences across the languages are all in the 45%-55% range. This filter could presumably be improved by requiring a shared named entity, but since language-independent NER is still an open research issue, we did not follow up on this avenue. We randomly sampled 1,000 of the remaining sentence pairs per language for manual annotation.

4.2 Distribution of annotation categories

First, we compared the frequencies of the annotation categories defined in Section 3.1. The results are shown in Table 2. We find our simple preprocessing filter results in an accuracy of between 75 and 82%. This is still considerably below the results of Hickl et al., who report 92% accuracy on their English data.[5]

Even though the overall percentage of "yes" cases is quite similar among languages, the details of the distribution differ. One fairly surprising observation was the fairly large number of ill-formed sentence pairs. As described in Section 2, this category comprises cases where the Hypothesis (i.e., a title) is not a grammatical sentence. Further analysis of the category shows that the common patterns are participle constructions (Ex. (6)) and combinations of multiple statements (Ex. (7)). The participle construction is particularly prominent in German.

(6) Glencoe Electric, Minn., rated single-A by Moody's.

(7) Wieder Kämpfe in Südlibanon - Israeli getötet.
Again fights in Southern Lebanon - Israeli killed.

The "no"-categories make up a total of 11.3% (English), 6.6% (German), and 20.7% (Hindi). The "other" and "partial" categories clearly dominate. This is to be expected, in particular the high number of partial entailments. The "other" category mostly consists of cases where the title summarizes the whole article, but the first sentence provides only a gentle introduction to the topic:

(8) **P:** One automotive industry analyst has dubbed it the 'Lincoln Town Truck'.
H: Ford hopes Navigator will lure young buyers to Lincoln.

As regards the high ratio of "no-other" cases in the Hindi corpus, we found a high number of instances where the title states the gist of the article too differently from the first sentence to preserve entailment:

[4] http://www.webdunia.com

[5] We attribute the difference to the filtering scheme which is difficult to reconstruct from Hickl et al. (2006).

Predictor	German	sig	English	sig	Hindi	sig
weighted overlap	0.77	**	2.30	***	3.35	***
log num words	-0.05	–	0.03	–	-0.17	–
punctuation	-1.04	***	-0.43	**	-0.35	**
strict noun match	0.12	–	0.19	–	0.38	**

Table 3: Exp. 1: Predictors in the logreg model (*: p<0.05; **: p<0.01; ***: p<0.001)

(9) **P:** आज भी प्रिंसेस डायना की लोकप्रियता कम नहीं हुई है ।
Even today, Princess Diana's popularity has not decreased.
H: प्रिंसेस डायना के पत्र और कार्ड नीलाम होंगे ।
Bidding on Princess Diana's letter and cards would take place.

The remaining error categories (embedding, presupposition, contradiction) were, disappointingly, almost absent. Another sizable category is formed by errors, though. We find the highest percentage for English, where our sentence splitter misinterpreted full stops in abbreviations as sentence boundaries.

4.3 Modelling the data

We estimated logistic regression models on each dataset, using the predictors from Section 3.1. Considering the eventual goal of extracting entailment pairs, we use the decision yes vs. everything else as our response variable. The analysis was performed with `R`, using the `rms`[6] and `ROCR`[7] packages.

Analysis of predictors. The coefficients for the predictors and their significances are shown in Table 3. There is considerable parallelism between the languages. In all three languages, weighted overlap between H and P is a significant predictor: high overlap indicates entailment, and vice versa. Its effect size is large as well: Perfect overlap increases the probability of entailment for German by a factor of $e^{0.77} = 2.16$, for English by 10, and for Hindi even by 28. Similarly, the punctuation predictor comes out as a significant negative effect for all three languages, presumably by identifying ill-formed sentence pairs. In contrast, the length of the article (log num words) is not a significant predictor. This is a surprising result, given our hypothesis that long articles often involve an "introduction" which reduces the chance for entailment between the title and the first sentence. The explanation is that the two predictors, log num words and weighted overlap, are highly significantly correlated in all three corpora. Since weighted overlap is the predictive of the two, the model discards article length.

Finally, strict noun match, which requires that all nouns match between H and P, is assigned a positive coefficient for each language, but only reaches significance for Hindi. This is the only genuine cross-lingual difference: In our Hindi corpus, the titles are copied more verbatim from the text than for English and German (median weighted overlap: Hindi 0.76, English 0.72, German 0.69). Consequently, in English and German the filter discards too many entailment instances. For all three languages, though, the coefficient is small – for Hindi, where it is largest, it increases the odds by a factor of $e^{0.39} \approx 1.4$.

Evaluation. We trained models on the three corpora, using only the two predictors that contributed significantly in all languages (weighted overlap and punctuation), in order to avoid overfitting on the individual datasets.[8] We applied each model to each dataset. How such models should be evaluated depends on the intended purpose of the classification. We assume that it is fairly easy to obtain large corpora of newspaper text, which makes precision an issue rather than recall. The logistic regression classifier assigns a probability to each datapoint, so we can trade off recall and precision. We fix recall at a reasonable value (30%) and compare precision values.

[6] `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/Design`
[7] `http://rocr.bioinf.mpi-sb.mpg.de/`
[8] Subsequent analysis of "full" models (with all features) showed that they did not generally improve over two-feature models.

Data \ Models	German model	English model	Hindi model
German data	**91.6**	88.8	88.8
English data	93.2	94.3	**94.6**
Hindi data	98.7	98.7	**99.1**

Table 4: Exp. 1: Precision for the class "yes" (entailment) at 30% Recall

Our expectation is that each model will perform best on its own corpus (since this is basically the training data), and worse on the other languages. The size of the drop for the other languages reflects the differences between the corpora as well as the degree of overfitting models show to their training data.

The actual results are shown in Table 4.3. The precision is fairly high, generally over 90%, and well above the baseline percentage of entailment pairs. The German data is modelled best by the German model, with the two other models performing 3 percent worse. The situation is similar, although less pronounced, on Hindi data, where the Hindi-trained model is 0.4% better than the two other models. For English, the Hindi model even outperforms the English model by 0.3%[9], which in turn works about 1% better than the German model. In sum, the logistic regression models can be applied very well across languages, with little loss in precision. The German data with its high ratio of ill-formed headlines (cf. Table 2) is most difficult to model. Hindi is simplest, due to the tendency of title and first sentence to be almost identical (cf. the large weight for the overlap predictor).

5 Experiment 2: Analysis by Domain of German corpora

5.1 Data

This experiment compares three German corpora from different newspapers to study the impact of domain differences: Reuters, "Stuttgarter Zeitung", and "Die Zeit". These corpora differ in domain and in style. The Reuters corpus was already described in Section 4.1. "Stuttgarter Zeitung" (StuttZ) is a daily regional newspaper which covers international business and politics like Reuters, but does not draw its material completely from large news agencies and gives more importance to regional and local events. Its style is therefore less consistent. Our corpus covers some 80,000 sentences of text from StuttZ. The third corpus comprises over 4 million sentences of text from "Die Zeit", a major German national weekly. The text is predominantly from the 2000s, plus selected articles from the 1940s through 1990s. "Die Zeit" focuses on op-ed pieces and general discussions of political and social issues. It also covers arts and science, which the two other newspapers rarely do.

5.2 Distribution of annotation categories

We extracted and annotated entailment pair candidates in the same manner as before (cf. Section 4.1). The new breakdown of annotation categories in Table (10) shows, in comparison to the cross-lingual results in Table 2, a higher incidence of errors, which we attribute to formatting problems of these corpora. Compared to the German Reuters corpus we considered in Exp. 1, StuttZ and Die Zeit contain considerably fewer entailment pairs, most notably Die Zeit, where the percentage of entailment pairs is just 21.6% in our sample, compared to 82.3% for Reuters. Notably, there are almost no cases where the first sentence represents a partial entailment; in contrast, for more than one third of the examples (33.9%), there is no entailment relation between the title and the first sentence. This seems to be a domain-dependent, or even stylistic, effect: in "Die Zeit", titles are often designed solely as "bait" to interest readers in the article:

(10) **P:** Sat.1 sah [...] Doris dabei zu , wie sie [...] Auto fahren lernte.
Sat.1 watched [...] Doris , how she [...] to drive learned.

[9]The English model outperforms the Hindi model at higher recall levels, though.

Corpus	err	ill	no-con	no-emb	no-oth	no-par	no-pre	yes
Reuters	3.5	2.9	0	0.2	3.7	7.4	0	82.3
StuttZ	6.2	3.6	0.5	2.8	12.4	3.0	0.6	70.7
Die Zeit	2.3	39.0	0.5	1.8	33.9	0.9	0.0	21.6

Table 5: Exp. 2: Distribution of annotation categories on German corpora (in percent)

Predictor	Reuters	sig	StuttZ	sig	Die Zeit	sig
weighted overlap	0.77	**	1.82	***	2.60	***
log num words	-0.05	–	-0.24	–	-0.20	–
punctuation	-1.04	***	-0.01	–	-1.21	***
strict noun match	0.12	–	0.20	–	-0.01	–

Table 6: Exp. 2: Predictors in the logreg model (*: p<0.05; **: p<0.01; ***: p<0.001)

Data \ Models	Reuters	StuttZ	Die Zeit
Reuters	**91.6**	85.4	**91.6**
StuttZ	**83.0**	**83.0**	82.6
Die Zeit	45.2	45.2	**46.7**

Table 7: Exp. 2: Precision for the class "yes" at 30% recall

H: Doris, es ist grün!
Doris, it is green!

Other titles are just noun or verb phrases, which accounts for the large number (39%) of ill-formed pairs.

5.3 Modelling the data

Predictors and evaluation. The predictors of the logistic regression models for the three German corpora are shown in Table 6. The picture is strikingly similar to the results of Exp. 1 (Table 3): weighted overlap and punctuation are highly significant predictors for all three corpora (except punctuation, which is insignificant for StuttZ); even the effect sizes are roughly similar. Again, neither sentence length nor strict noun match are significant. This indicates that the predictors we have identified work fairly robustly. Unfortunately, this does not imply that they always work well. Table 6 shows the precision of the predictors in Exp. 2, again at 30% Recall. Here, the difference to Exp. 1 (Table 4.3) is striking. First, overfitting of the predictors is worse across domains, with losses of 5% on Reuters and Die Zeit when they are classified with models trained on other corpora even though use just two generic features. Second, and more seriously, it is much more difficult to extract entailment pairs from the Stuttgarter Zeitung corpus and, especially, the Die Zeit corpus. For the latter, we can obtain a precision of at most 46.7%, compared to >90% in Exp. 1.

We interpret this result as evidence that domain adaptation may be an even greater challenge than multilinguality in the acquisition of entailment pairs. More specifically, our impression is that the heuristic of pairing title and first sentence works fairly well for a particular segment of newswire text, but not otherwise. This segment consists of factual, "no-nonsense" articles provided by large news agencies such as Reuters, which tend to be simple in their discourse structure and have an informative title. In domains where articles become longer, and the intent to entertain becomes more pertinent (as for Die Zeit), the heuristic fails very frequently. Note that the weighted overlap predictor cannot recover all negative cases. Example (10) is a case in point: one of the two informative words in H, "Doris" and "grün", is in fact in P.

Domain specificity. The fact that it is difficult to extract entailment pairs from some corpora is serious exactly because, according to our intuition, the "easier" news agency corpora (like Reuters) are domain-

Corpus	D($\cdot$ | deWac)	words w with highest $P(w)/Q(w)$
Reuters	0.98	Händler (trader), Börse (exchange), Prozent (per cent), erklärte (stated)
StuttZ	0.93	DM (German Mark), Prozent (per cent), Millionen (millions), Geschäftsjahr (fiscal year), Milliarden (billions)
Die Zeit	0.64	heißt (means), weiß (knows), läßt (leaves/lets)

Table 8: Exp. 2: Domain specificity (KL distance from deWac); typical content words

specific. We quantify this intuition with an approach by Ciaramita and Baroni (2006), who propose to model the representativeness of web-crawled corpora as the KL divergence between their Laplace-smoothed unigram distribution P and that of a reference corpus, Q ($w \in W$ are vocabulary words):

$$D(P,Q) = \sum_{w \in W} P(w) \log \frac{P(w)}{Q(w)} \tag{4}$$

We use the deWac German web corpus (Baroni et al., 2009) as reference, making the idealizing assumption that it is representative for the German language. We interpret a large distance from deWac as domain specificity. The results in Table 8 bear out our hypothesis: Die Zeit is less domain specific than StuttZ, which in turn is less specific than Reuters. The table also lists the content words (nouns/verbs) that are most typical for each corpus, i.e., which have the highest value of $P(w)/Q(w)$. The lists bolster the interpretation that Reuters and StuttZ concentrate on the economical domain, while the typical terms of Die Zeit show an argumentative style, but no obvious domain bias. In sum, domain specificity is inversely correlated with the difficulty of extracting entailment pairs: from a representativity standpoint, we should draw entailment pairs from Die Zeit.

6 Conclusion

In this paper, we have discussed the robustness of extracting entailment pairs from the title and first sentence of newspaper articles. We have proposed a logistic regression model and have analysed its performance on two datasets that we have created: a cross-lingual one a cross-domain one. Our cross-lingual experiment shows a positive result: despite differences in the distribution of annotation categories across domains and languages, the predictors of all logistic regression models look remarkably similar. In particular, we have found two predictors which are correlated significantly with entailment across (almost) all languages and domains. These are (a), a tf/idf measure of word overlap between the title and the first sentence; and (b), the presence of punctuation indicating that the title is not a single grammatical sentence. These predictors extract entailment pairs from newswire text at a precision of $> 90\%$, at a recall of 30%, and represent a simple, cross-lingually robust method for entailment pair acquisition.

The cross-domain experiment, however, forces us to qualify this positive result. On two other German corpora from different newspapers, we see a substantial degradation of the model's performance. It may seem surprising that cross-domain robustness is a larger problem than cross-lingual robustness. Our interpretation is that the limiting factor is the degree to which the underlying assumption, namely that first sentence entails the title, is true. If the assumption is true only for a minority of sentences, our predictors cannot save the day. This assumption holds well in the Reuters corpora, but less so for the other newspapers. Unfortunately, we also found that the Reuters corpora are at the same time thematically constrained, and therefore only of limited use for extracting a representative corpus of entailment pairs. A second problem is that the addition of features we considered beyond the two mentioned above threatens to degrade the classifier due to overfitting, at least across domains.

Given these limitation of the present headline-based approach, other approaches that are more generally applicable may need to be explored. Entailment pairs have for example been extracted from Wikipedia (Bos et al., 2009). Another direction is to build on methods to extract paraphrases from comparable corpora (Barzilay and Lee, 2003), and extend them to capture asymmetrical pairs, where entailment holds in one, but not the other, direction.

Acknowledgments. The first author would like to acknowledge the support of a WISE scholarship granted by DAAD (German Academic Exchange Service).

References

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation 43*(3), 209–226.

Barzilay, R. and L. Lee (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*, Edmonton, AL, pp. 16–23.

Bos, J., M. Pennacchiotti, and F. M. Zanzotto (2009). Textual entailment at EVALITA 2009. In *Proceedings of IAAI*, Reggio Emilia.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation*, pp. 69–94. Royal Netherlands Academy of Science.

Burger, J. and L. Ferro (2005). Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp. 49–54.

Ciaramita, M. and M. Baroni (2006). A figure of merit for the evaluation of web-corpus randomness. In *Proceedings of EACL*, Trento, Italy, pp. 217–224.

Dagan, I., O. Glickman, and B. Magnini (2006). The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, Volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer.

de Marneffe, M.-C., A. N. Rafferty, and C. D. Manning (2008). Finding contradictions in text. In *Proceedings of the ACL*, Columbus, Ohio, pp. 1039–1047.

Hickl, A., J. Williams, J. Bensley, K. Roberts, B. Rink, and Y. Shi (2006). Recognizing textual entailment with LCC's Groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop*.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval* (1st ed.). Cambridge University Press.

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

Sammons, M., V. Vydiswaran, and D. Roth (2010). "Ask Not What Textual Entailment Can Do for You...". In *Proceedings of ACL*, Uppsala, Sweden, pp. 1199–1208.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the SIGDAT Workshop at ACL*, Cambridge, MA.

Xiao, Z., T. McEnery, P. Baker, and A. Hardie (2004). Developing Asian language corpora: Standards and practice. In *In Proceedings of the Fourth Workshop on Asian Language Resources*, Sanya, China, pp. 1–8.