Cross-Corpus Textual Entailement for Sublanguage Analysis in Epidemic Intelligence

Avaré Stewart , Kerstin Denecke , Wolfgang Nejdl

L3S Research Center
Appelstr. 9A, 30169 Hannover, Germany
stewart@L3S.de, denecke@L3S.de, nejdl@L3S.de

Abstract

Textual entailment has been recognized as a generic task that captures major semantic inference needs across many natural language processing applications. To date, textual entailment has not been considered in a cross-corpus setting, nor for user generated content. The emergence of Medicine 2.0, has made medical blogs an increasingly accepted source of information; but given the characteristics of blogs (which tend to be noisy and informal; or contain a interspersing of subjective and factual sentences) a potentially large amount of irrelevant information may be present. Considering this potential noise, the overarching problem with respect to information extraction from social media for medical intelligence gathering, is achieving the correct level of sentence filtering - as opposed to document or blog post level. In this paper, we propose an approach to textual entailment which uses the text from one source of user generated content (T text) for sentence-level filtering within a new and less amenable one (H text), when the underlying domain, tasks or semantic information is the same, or overlaps.

1. Introduction

Textual entailment has been recognized as a generic task that captures major semantics inference needs across many natural language processing applications (Iftene, 2009). However, to date textual entailment has not been considered in a cross-corpus setting for medical blogs nor user generated content.

Given the emergence of Medicine 2.0, social medical blogs are becoming an increasingly accepted source of information for public health event extraction. For example, consider a typical *Medical Intelligence Gathering* scenario (see Figure 1) where disease-reporting events (i.e. victim, location, time, disease) are extracted from raw text to get information about potential disease outbreaks. The events are then aggregated to produce signals; which are intended to be an early warning against potential public health threats. Epidemiologists use them to assess risk, or corroborate and verify the information locally and with international agencies.

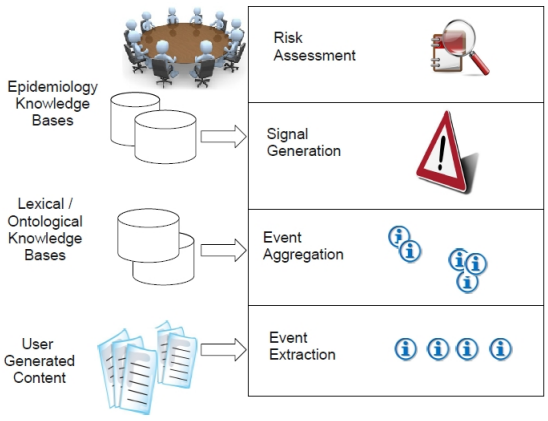

Figure 1: Epidemic Intelligence Scenario

There are challenges faced in extracting events in such a scenario. Given the nature of blog data, acquiring information relevant to public health events requires filtering a huge number of irrelevant events, some based on hypothetical events or opinions (Denecke and Nejdl, 2009).

Though abundant, a rich set of annotated data for linguistic tasks is scarce, or non-existent. Moreover, given the dynamic nature of social media, potentially may types of event extraction patterns may be needed, even for extracting a single type of event. Furthermore, blogs are characteristically noisy and informal, containing a interspersing of subjective and factual sentences, or even many authors with different styles. Given the potential noise , the overarching problem with respect to the event extraction from social media is achieving the correct level of sentence filtering - as opposed to document or blog post level. Finally, for the intelligence gathering task, today's noise can be tomorrow's information, so adaptable filtering is desired.

In this paper we propose an approach to textual entailment with uses the text from one source of user generated content (T text) for sentence-level filtering within a new and less amenable one (H text), when the underlying domain, tasks or semantic information is the same, or overlaps.

The contribution of this work are three-fold: (1) an exploratory data analysis on the sub-language of medical blog content for epidemic intelligence gathering; (2) overlay approach to cross-corpus textual entailment; and (3) the use of cross-corpus textual entailment as a pre-processing step in the linguistic pipeline for adaptive sentence level filtering

2. Related Work

The related work in this area present approaches which rely upon a global (instead of a local) view for sub- and related tasks in event extraction. Numerous tasks , as well as approaches are used for defining an overlap and matchmaking across document spaces.

2.1. Inter-Document Analysis

Approaches which make use of large amounts of unannotated text and inter-document information have emerged as an important approach for dynamic enrichment when the context within a single document is absent or incomplete. In recent work, an inter-document approach is used for building a consistent view for the sense of a word or the role of an argument (Ji and Grishman, 2008). In other work on cross-document event extraction, the authors identify relations between events based on them having shared arguments (Heng Ji and Gupta., 2009). They present new tasks and metrics for the cross-document setting to gauge the effectiveness of a cross-document IE system. Their approach is measured based on how well a system performs in selecting the correct centroid entities in a set of documents. In contrast to their work, the measure we introduce is used, instead, to judge the quality of the overlapping sentences across the corpora.

2.2. Textual Entailment

Textual entailment recognition is the task of deciding, given two text fragments, whether the meaning of one text is entailed (or inferred) from another text (Iftene, 2009). A similar approach to alignment is presented in this work (Y. Mehdad, 2009). The authors use text-based similarity and consider an overlap at the term level. Their measure for word overlap is based on the number of words shared between T and H. Differently however, we consider levels of abstraction for representing a sentence that goes beyond the term level.

Finally, related work in the area of event extraction for epidemic intelligence which considers outbreak reports as a form of user generated content such as ((Yangarber, 2006; Yangarber and Jokipii, 2005)) do not consider the use of textual entailment across corpora for pre-processing and filtering.

3. Cross-Corpus Pattern Analysis

Cross-Corpus Textual Entailment can be seen as a type of pattern matching where the patterns sought do not come from a single corpus; but are instead discovered across the corpora and rules can be learnt to infer an aligned linguistic pattern between a text (T) and the hypothesis (H) [1].

Let $D = d_1, d_2, \ldots, d_j$ be the set of documents in a corpus, where each document, $d_i \in D$, consists of a set of one or more sentences. Further, let $I_t = i_{t1}, i_{t2}, ..., i_{tm}$ be the set of literals for a linguistic annotation type t, applicable to the words in the sentences of D.

We define a non-empty subset, $X \subseteq I_t$, to be an itemset. Then, the process of annotating one or more words in a sentences of D induces a **sequence**, or an ordered list of itemsets, represented as: $s_t =< X_{t1}, X_{t2}, ..., X_{tl} >$, where $l = |s_t|$ for a given sentence and annotation type. The elements in s_t are in lexicographic order as defined by the words in the given sentence.

Sentence Sequence Database

Given a set of documents D and linguistic annotation type I_t, applied to the sentences in D, a corpus can be modeled as a sentence sequence database, $S_t = s_{t11}, s_{t12}, ..., s_{tjk}$, where s_{tjk} represents the k^th sentence for the j^th document using linguistic annotation type t.

Given a Sentence Sequence Database S_a for an auxiliary domain a, and a second Sentence Sequence Database, S_n for a target domain n, and $M(S_a)$ and $M(S_n)$ a set of sequential patterns taken from the Sentence Sequence Database for an auxiliary domain and target domain, respectively. Then, Cross-Corpus pattern analysis can be defined as:

Cross-Corpus Pattern Analysis: Finding aligned subsequences of the form $X_m - > Y_n$, where Y_n is a sequence from target domain, X_a is a sequence from auxiliary domain and X_a is produced from Y_n by replacing zero or more of its items with wildcards. If zero replacements are made, then sequences are simply taken to be identical across the domains.

Sub-Language Sentence Weighting: Further, given a sentence sequence database, we weight each sentence to distinguish them as being most relevant for a particular task. In this work, we focus on the disease reporting task and define a set of entity types, $E = Disease, Date, GeographicLocation, Person$, which are indicative of disease reporting, and compute a weight w for each sentence as follows:

$$weight(sentence) = \begin{cases} \sum_{e \in E} w_e \left[\frac{1}{Count_e(sentence)} \right] \\ 0, otherwise \end{cases}$$

We can adjust the weights of the selected entity types; for example $w_{disease}$ = .7 and $w_{geographiclocation}$, w_{person}, w_{date} = .1. This will allow us to assign more weight to sequences containing mentions of diseases and parameterize the level of noise to filter. The scheme incorporates several assumptions: First, entity types which occur once or twice in a sentence are considered to be a better indicator of disease reporting. Since multiple occurrences of the same entity type in a sequence is considered more likely to be noise, we desire a score in which a higher weight is reached when there is a single occurrence of each entity type in a sequence (effectively increasing the weight of terms that occur rarely). Finally, the Sentence Weight value is bounded by 0...1, inclusive.

4. Experiments

In this section, preliminary results of an exploratory data analysis in cross-corpus analysis is presented.

4.1. Data Sets

ProMED-Mail outbreak reports and medical blogs are used as data sets. The ProMED data was collected directly from the website, and MedWorm medical blogs via summarized RSS. The blog data was collected by retrieving blogs from one hundred randomly selected categories under the heading of *infectious disease*. The data for the outbreak reports was collected during an eight month period from January 1, 2009 through August 12, 2009; and for blogs, from May

[1] http://pascallin.ecs.soton.ac.uk/Challenges/RTE/

11, 2009 through August 13, 2009. In total, 14,816 MedWorm and 1,120 ProMED documents were used in the experiments. Documents below a size of 300 Kb, as well as raw sentences below a length of 51 characters, were filtered out. These sentences were mostly due to the incomplete sentences or noise present in the data. In total 49,711 MedWorm and 26,635 ProMED sequences were used.

Sentence Sequence Database For constructing the Sentence Sequence Databases, two types of annotations were used. The first annotation was based on concepts present in the UMLS lexicon (UMLS CUI). The other was based on the semantic class of these concepts (UMLS TUI). In total, 22,256 distinct CUI literals were exploited. A collapsed version of the TUI candidate given as output by the annotation process was used. This collapsed version aggregates similar classes into a single class based on the $is-a$ relationship within the UMLS semantic network. This yielded a total 133 distinct TUI literals.

4.2. Results

Data Summarization

In Figure 2, the sizes of the sentence sequence database against the sequence length is presented. It can be noticed that there is a slight increase of the amount of sequences that have a length five. With increasing sequence length, the sequence database size drops greatly.

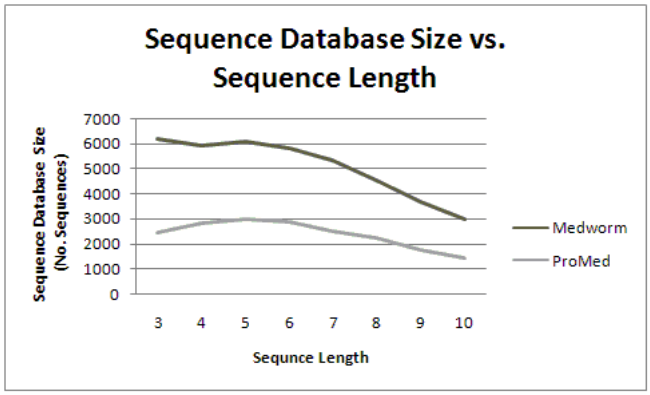

Figure 2: Sequence Database Sizes

Figure 3 shows the average sentence weight for sequences of varying length. As expected, we notice that the TUI annotations (semantic class level) annotation has higher average due to the fact that they represent an aggregation of the concept level. However (corresponding to the results in Figure 2) the TUI annotations peak at a sequence length of 5 and fluctuate thereafter. The overall low weights are quite noticeable and can be explained by the results shown in Figure 4.

4.2.1. Sequence Alignment

In Figure 4 we plot the cumulative relative frequency of the sentence weight to show the proportion of weights in the overlapping data for each annotation type that are less than or equal to a particular values. We notice that 80 percent of the overlapping sequences have a sentence weight of .35 or less. This can be explained by the choice of our sentence weight scheme, which effectively filters non-relevant sentences based on criteria of disease reporting.

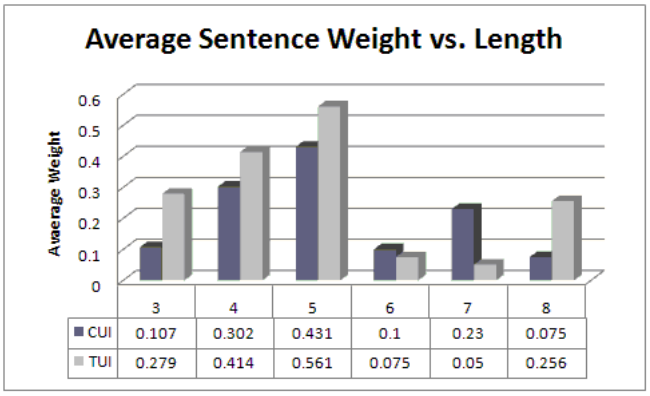

Figure 3: Average Sentence Weight

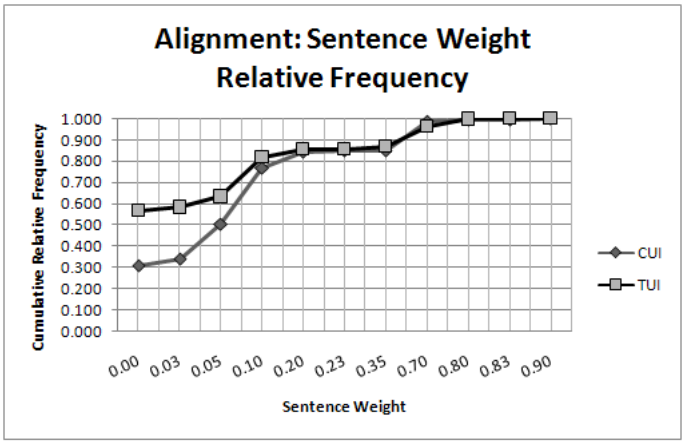

Figure 4: Cumulative Relative Frequency vs. Sentence Weight

4.2.2. Cross-Corpus Sequence Alignment

The sentences that overlapped across the copora for a sentence weight of .8 or higher fell into three categories - each category. In first category, the sentences were identical. In the second category, there was a slight variation is a single token. One example of this token-level difference is represented in Table 1; where the H and T texts differed only by the term “first” and “1st“.

In the third type of variation, we noticed the overall position of the entities were preserved, even though different words were present. This third type of category is represented in the H and T text shown in Table 2:

4.2.3. Assessment of Aligned Sequences

Finally, the quality of the aligned sequences is considered, we present two examples of the favorable and unfavorable sentence quality.

In Table 3, sentences which were weighted with a score of zero, but are actually relevant to disease reporting are shown. These results can be explained by the choice of our weighting scheme which considers multiple instances of a

ProMed Sentence (H text): *This report describes the first recognized case of imported simian malaria in several decades in the LOCATION, diagnosed in 2008 in a patient from LOCATION who had traveled to the LOCATION.* ***(Sentence Weight = .9)***

Medworm Sentence (T text): *This report describes the 1st recognized case of imported simian malaria in several decades in the LOCATION, diagnosed in 2008 in a patient from LOCATION who had traveled to the LOCATION.* ***(Sentence Weight = .9)***

Table 1: Sentence alignment with a single token variation between T and H texts.

ProMed Sentence (H text): *This was the 5th case of rabies reported in LOCATION this year.* ***(Sentence Weight = .9)***

MedWorm Sentence (T text): *There are approximately 40,000 cases of infection reported in the LOCATION each year.* ***(Sentence Weight = .9)***

Table 2: Sentence alignment in which the overall position of entities types are preserved.

single entity type to be noise, this includes comma separated list.

Ex 1: *They were adenovirus, 4.8 % , influenza virus , 1.0 % , RS virus , 6.3%, and human metapneumovirus , 2.9 %* ***(Sentence Weight = 0)***

Ex 2: *fatigue, anxiety, headache* ***(Sentence Weight = 0)***

Table 3: False Negative Example

4.3. Discussion

The Disease Reporting Weighting Scheme can be used to for task-specific adaptive filtering - achieving the desired information based on a task. A smoother weighting scheme should be considered to avoid steps in the data and to have a more fine-grained tuning over the filtering process. The weighting scheme could be extended to include a domain-independent component to achieve this smoothing or using a logarithm instead of an inverse measure. Finally, a classifier could also be learnt to handle the weighting and relevance for sentences. Yet another limitations in the weighting scheme is that other entity types, besides the selected ones, can have the role in classifying a sentence as disease reporting; for example, the class of medical findings, such as: “discomfort”, “sleepiness“.

Annotation Errors: In such an experimental setting, the results are only as good as the precision of the underlying annotation system. We notice that there are ambiguities (false positive and false negative) in the named entity system. Consider the example in Table 4, where the term ”Flu” is taken to be a unit.

Sentence: The real name for swine flu is H1N1 flu.
Meta Mapping (888):
694 C1963039:Swine [Immunologic Factor]
861 C2348686:FlU (Fluorescence Units) [Quantitative Concept]

Table 4: Example UMLS MetaMap Annotation Error

Moreover, the domain specific annotator tends to over emphasize the presence of biomedical concepts due to the high prevalence of medical abbreviations and acronyms. For example, the terms “ate” (past tense of eat) and “Oct.” (month of the year) are mapped to proteins. We have only made a work-around for these domain-specific over-specification errors by first preprocessing the documents with a domain independent annotator which replaced the names of persons, locations and organizations with a predefined keyword.Although an attempt was made to overcome the high prevalence of domain-specific error, the potential arises to introduce more error, that is derived instead, from the domain independent named entity recognition (NER). The results here would also suggest the need for a NER tagger to be trained on a medical corpus, so that CWD (Cronic Wasting Disease), would not be tagged as an organization.

5. Conclusion

In this early work we outlined an exploratory data analysis using Cross-Corpus Textual Entailment for Medical Intelligence gathering. The goal has been to consider an unsupervised techniques to automatically filtering disease reporting event extraction patterns at the sentence level in this more noisy setting. We have shown that the approach taken here is quite effective at sentence-level filtering in medical blogs. The sentence weighting scheme we used, also allow an adaptive level of filtering and allow overlapping sentences to be quantified in terms of their quality.

In this early work on cross-corpus analysis many interesting and new issue arise. So far, we have examined the overlap of blog documents with a single source. The question is whether the results presented here are also indicative of other sources produced by other online public health reporting systems such as World Health Organization. We will

consider how state of the art algorithms for textual entailment perform in a cross-corpus setting as well and test the approach presented here on other data sets. Finally, higher levels of abstractions based on dependency and syntatic parse trees will be taken into account; and a learner for overlapping patterns and the interchange of auxiliary and target domains will be investigated.

6. References

Kerstin Denecke and Wolfgang Nejdl. 2009. How valuable is medical social media data? content analysis of the medical web. *Inf. Sci.*, 179(12):1870–1880.

Zheng Chen Heng Ji, Ralph Grishman and Prashant Gupta. 2009. Cross-document event extraction and tracking: Task, evaluation, techniques and challenge. *Proc. Recent Advances in Natural Language Processing*.

Adrian Iftene. 2009. *Textual Entailment*. Phd-thesis, University of Iasi.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through unsupervised cross-document inference. *In Proceedings of the Annual Meeting of the Association of Computational Linguistics*.

B. Magnini. Y. Mehdad. 2009. A word overlap baseline for the recognizing textual entailment task. *Proc. Recent Advances in Natural Language Processing*.

Roman Yangarber and Lauri Jokipii. 2005. Redundancy-based correction of automatically extracted facts. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

Roman Yangarber. 2006. Verification of facts across document boundaries. *In Proceedings International Workshop on Intelligent Information Access*.