Assessing the Comparability of News Texts

Emma Barker and Rob Gaizauskas

Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, UK
E.Barker,R.Gaizauskas@sheffield.ac.uk

Abstract

Comparable news texts are frequently proposed as a potential source of alignable sub-sentential fragments for use in statistical machine translation systems. But can we assess just how potentially useful they will be? In this paper we first discuss a scheme for classifying news text pairs according to the degree of relatedness of the events they report and investigate how robust this classification scheme is via a multi-lingual annotation exercise. We then propose an annotation methodology, similar to that used in summarization evaluation, to allow us to identify and quantify shared content at the sub-sentential level in news text pairs and report a preliminary exercise to assess this method. We conclude by discussing how this works fits into a broader programme of assessing the potential utility of comparable news texts for extracting paraphrases/translational equivalents for use in language processing applications.

Keywords: Comparable Corpora, Machine Translation, Paraphrase Acquisition, News Discourse, Evaluation Methodologies

1. Introduction

The idea that comparable corpora have the potential to be important resources for a variety of language processing tasks that require examples of different linguistic realizations of semantically equivalent content has been around for some time. Statistical machine translation (SMT), paraphrase acquisition and textual entailment are all examples of tasks for which efforts have been made to exploit comparable corpora.

In the case of SMT, the argument for comparable corpora goes as follows. SMT systems require very large amounts of parallel data to achieve reasonable levels of translation quality. However, the number of parallel documents available for less widely spoken languages is in general not sufficient to achieve acceptable translation performance. Thus, various researchers have turned to *bilingual comparable corpora* – collections of texts in two languages which are similar in content while not being direct translations – with the hope that alignable words, fragments or sentences may be extracted from them and used in training SMT systems.

The notion of comparability – like that of similarity in terms of which it frequently explained – is a difficult one. Two things may be comparable or similar in one or more aspects, but not in others. Text pairs may be comparable because they share propositional content, i.e. say the same things about the same entities (e.g. two news reports describing the same event) or because they are more loosely topically similar (e.g. two texts on greenhouse gas emissions) or are drawn from the same domain (e.g. medicine); or they may exhibit similarity in genre (e.g new texts vs scientific papers). Each of these types of comparability has the potential to be exploited for SMT and other language processing tasks; however, the one that appears to offer the most promise is that of shared propositional content ("shared content" hereafter).

There has been considerable attention paid to the problem of how to extract semantically equivalent strings from comparable corpora, but the characteristics of the corpora themselves remain poorly understood. Their utility, and their inferred degree of correspondence, is typically measured via the performance improvement of some system which has exploited the shared content. While such results are useful in that they allow us to assess the success achieved for a particular application task, they only provide feedback on what exploitable shared content current techniques were able to find. They do not allow us to assess the total amount of potentially relevant content which may be contained in the corpora, and which the system was not able to find. To the best of our knowledge no comparable corpus with gold standard alignments of all shared content across comparable text pairs exists.

Questions which remain to be asked include: which types of comparable corpora are likely to give us a lot of shared content? are we likely to find many examples of full sentence equivalence? at what level and in what syntactic forms is semantically equivalent content likely to found?

1.1. Comparability and News Texts

One text type which occurs in virtually all languages in large volumes and in which one finds similar content being expressed across languages is news text. We live in a highly interconnected world and significant events taking place in any part of the world are likely to be reported in major newspapers everywhere at more or less the same time. While comment on events will differ everywhere, we expect basic factual reporting to convey the same message everywhere: Suu Kyi wins seat in election; a tsunami has hit North-East Japan, etc.

That news texts should be rich sources of shared content and hence are a type of comparable corpus of high potential value for SMT and other language processing tasks has of course not gone unnoticed. Various researchers have proposed techniques for gathering news stories about the same events and then for finding and extracting shared content in them for use in SMT systems. Munteanu and Marcu (2005), for example, propose a technique for finding parallel sentences within large corpora of comparable news texts and in Munteanu and Marcu (2006) they go further

and discuss techniques for extracting parallel sub-sentential fragments from comparable news texts. There has been a long tradition of work on bilingual lexicon extraction from comparable corpora ranging from Fung (1998) to Li et al. (2011), much of which uses news corpora. Comparable corpora have also been used in support of cross-language information retrieval (Braschler and Schäuble, 1998).

However, what these researchers have not done is: (1) offer an explanatory account of which news text pairs are likely to contain high levels of content overlap (which has potential to inform techniques for collecting them); (2) determine how to assess how much content overlap there is between two comparable news texts. (3) provide a gold standard resource in which all pairs of sentence bearing shared content, within a text pair, are identified. In this paper we address all of these questions.

In section 2 we offer a coarse-grained analysis of the functional structure of news texts on the basis of which we propose a scheme for classifying news text pairs into various classes reflecting the nature of their relatedness as reports of new events. We report experiments showing the intersubjective reliability of the scheme across multiple language pairs. In Section 3 we present a method for identifying and assessing shared content in two news texts. We also describe a preliminary experiment which we carried out to test the reliability of the method and to gain initial insights into how much shared content there is and what form it takes in comparable news texts. Our hypothesis is that the amount of shared content between two news texts will correlate strongly with the relatedness classes of Section 2. In Section 4 we conclude by describing how this work fits into a broader programme for assessing the potential utility of comparable news texts for applications such as machine translation.

2. Exploiting the Functional Structure of News Texts

The idea that two news stories on the same event are likely to have some content in common is intuitive and one which has informed previous work. The basic idea is that texts on the same event are talking about the same thing and therefore are likely to say similar things in their reports. However, if we take a closer look at how events are reported in news, both by examining real examples of news text and taking into account previous studies of news text, we find that the picture is not so straightforward. In particular, the notion of "same news event" is not at all clear and has been differently interpreted by different authors and left unanalyzed by others. However, the extent to which texts about the "same news event" share content depends critically on just what this notion means. For example, consider three texts about Obama's one day visit to Ireland in spring 2011 – one in the Irish Times, the Irish "newspaper of record" (*Obama hails strong ties between US and Ireland*), another in Newsletter.co.uk (*Obama hails Ulster peace on Republic visit*) and a third in Hello Magazine, a celebrity news magazine (*'Thrilled' Barack Obama returns to his Irish roots*). The angle on the story and consequently the content diverge considerably: would two people judge these to be on the "same event"?

If we start with the notion of a news event as something that happens which is of significance to a considerable number of people, where prototypical examples are events like earthquakes, elections, terrorist attacks, company takeovers, etc. then several typical characteristics of such events and their reporting may be noted: (1) news events are ongoing, evolving situations, typically reported in multiple texts over periods of time ranging from hours to weeks; (2) individual news event reports are complex, hierarchically structured discourses comprising multiple lower level events; (3) the focus or 'angle' taken on a news event depends on various factors including the current point in the evolution of the story and the perspective of the reporter or newspaper; (4) the practice of rewriting or editing previous copy, either the newspaper's own or that of a news agency to which the paper subscribes, is ubiquitous.

In this section we review relevant previous work, present our analysis of the functional structure of news texts and our scheme for classifying news text pairs based on this scheme and finally describe annotation experiments carried out to assess the intersubjective reliability of the scheme.

2.1. Previous Work

Various authors working the general area of discourse analysis have analyzed the structure of news articles. van Dijk (1985) presents an extended analysis within the framework of a more general theory of discourse. Bell (1998) provides a highly detailed account of the structure of news texts, in which a news story which may be broken down into episodes and where episodes consist of one or more events, the events being composed of attribution, actors, action, setting, follow-up, commentary, background and previous episodes (episodes are recursive and can appear under a number of categories). In addition Bell considered news as "narrative", by comparing Labov's 6 functional categories for narrative text (Labov and Waletzky, 1997) with news content. He showed that the categories 'Abstract', 'Action', 'Evaluation' and 'Orientation' were of particular relevance to news: the Abstract (the headline and lead) summarising the main events and establishing the point of the story; Orientation (the headline and story text) establishing the setting; and Evaluation (the lead and story text) providing the journalist's evaluation of the events and establishing their significance. Bell also noted how the lead or first paragraph focuses the story in a particular direction.

We differ in our objectives in that we are not interested in a fine-grained analysis of the news text genre but rather in elaborating a descriptive scheme that, taking into account certain features of news, will allow us to predict shared content across multiple news texts – not something these authors are interested in. The concept of focus in news text is of some relevance to our problem. We refine this idea, identifying the "focal events" as being those events which provide a focus for the text (see next section).

In this work on news discourse, the authors provide an analysis in which they treat a news story as co-extensive with the events reported in a single text. By contrast, within the computational language processing community, the idea of the "same news event" being the subject of multiple news reports has been investigated in various contexts.

Class	Sub Class	Possible shared content patterns
SAME NEWS_EVENT	SAME FOCAL EVENTS	Common focal event, common elaboration, common background and common quotes.
	DIFFERENT FOCAL EVENTS	Focal event in one text appears as background in the other, common background and common quotes.
DIFFERENT NEWS_EVENTS, SAME TYPE	FOCAL EVENTS SAME TYPE	Similar event structure for focal events, background in common (e.g. accounts of other similar events), summary of one text's news event appears as background in the other.
	FOCAL EVENTS DIFFERENT TYPE	Background in common; details from one text's news event appears as background in the other.
DIFFERENT NEWS_EVENTS, DIFFERENT TYPE	RELATED via BACKGROUND	Details from one text's news event appear as background in the other. Background in common.
	Other	No content in common

Table 1: Event Relatedness Classes

The Topic Detection and Tracking (TDT) challenge introduced a distinction between news topics, stories and events (Fiscus and Doddington, 2002). In their terminology a news topic is defined as "a seminal event or activity, along with all directly related events and activities" and corresponds more or less to what we referred to as the "everyday" notion of news event above. Core to TDT task is the presumption that multiple stories (individual written or spoken language news reports) over time will address the same topic – indeed one of the main subtasks is to link such stories together. However, our interest – how much content different stories on the same topic share – is not a question directly addressed.

Various authors working on paraphrase acquisition have exploited monolingual comparable corpora of news stories (Shinyama et al., 2002; Barzilay and Lee, 2003; Dolan et al., 2004). The corpora are created by using word-based clustering or retrieval techniques over news texts drawn from various sources. Dolan et al. claim that texts in the same cluster are "generally coherent in topic and focus", but note that certain on-going event types lead to more focussed clusters than others, acknowledging that there are varying degrees of similarity in news texts on the same topic. However, there is no further investigation of this observation or any attempt to assess how shared content, and hence perhaps paraphrase yield, relates to topic drift.

The only work we are aware of that attempts to categorize news text pairs in comparable corpora based on how related the events they describe are is Braschler and Schaüble (1998), who exploit comparable corpora in an approach to cross language retrieval. Since the quality of their retrieval results is highly dependent on the quality of the document alignment in the comparable corpus, the authors evaluate the alignment process. To do so they introduce a five class scheme for assessing the similarity of two news texts, with classes for *same story*, *related story*, *shared aspect*, *common terminology* and *unrelated* and ask human judges to assess document pairs aligned using their algorithm according to this scheme. The notions of *same story* and *related story* are not defined or analyzed further, just illustrated with an example. No results are quoted for human agreement on the task of identifying *same story* new pairs.

2.2. The Functional Structure of News Texts

To further investigate functional patterns in news we hand picked a small number of related stories from the on-line news domain.

Our development collection included: (1) texts on related events published at different points in time (e.g. reports of a volcanic eruption and its potentially hazardous ash cloud, warnings of the knock-on disruption to airlines, warnings of the ash cloud risks to public health, etc.); texts on events published at the same, or very similar points in time (e.g. early reports on an earthquake); and examples of similar, but different events at different points in time (e.g. reports on two different hurricanes).

We identified relations between events both within a text and between different texts. Key concepts are as follows:

We view an event as a specific thing that happens at a particular time and place.

Focal event: the event or events which provide a focus for the text. Very often the most recent event in an unfolding news story, they also provide a particular angle or perspective for the report. Typically reported in the headline and first few lines of a news report, we may find in the body text: a fuller account (i.e. elaboration) of the focal events; background to the focal events and details of possible or actual subsequent events.

Background event: an event that plays a supporting role in the text, providing context for the focal events. May include: related events leading up to the focal events; examples of similar past events; and definitions, explanations or descriptions of things, people and or places which play a role in the focal events.

NewsEvent: a group of related events, broader than and including the focal event, which may be reported over time in different news text instalments. E.g. initial reports on an earthquake NewsEvent include details of a quake having occurred, while later reports cover rescue attempts, accounts of disaster aid and relief, etc. In such a case we view the texts as reporting on the same NewsEvent. Note in later reports, background events may include details of previous events in the NewsEvent.

Quotes: reported speech, typically indicated by quotation marks. May be part of the fuller account of the focal events or be part of the background.

Informed by this coarse-grained analysis of the functional structure of news texts, we developed a scheme for classifying pairs of news texts, where the classes are indicative of the relation holding between the events reported in the texts. A summary of the scheme is shown in Table 1, together with the types of shared content we can expect to see.

2.3. Investigating the Scheme

2.3.1. Method

To investigate the level of inter-subjective agreement obtainable between subjects asked to classify text pairs according to this scheme, we developed a web-based interface, via which we asked participants to assess collections of text pairs, based on the categories described above. We ran two pilot exercises, one using 50 English text pairs and one using 8 sets of 10 text pairs, where each contained pairs consisting of one English text and one text in of 8 other European languages being studied in the Accurat project[1]. In each case the text pairs were gathered using a tool we developed for harvesting comparable news text pairs from the Web (see Aker et al. (2012) for details). Because content on news sites does not reliably remain available we downloaded the texts, extracted the text content, by removing boilerplate and advertising and stripping out the html, and then re-introduced some light html formatting. We ran language id filters to discard texts not in the language we were trying to collect. These pilot exercises let us refine our interface and our guidelines for describing our news event relatedness scheme and assured us that agreement was solid enough to warrant running a larger experiment.

In the full experiment we again used our comparable news retrieval tool to collect sets of candidate comparable news text pairs. For each text pair collected the tool produces a score which indicates its assessment of the comparability of the two texts. The tool has a threshold below which it does not consider the pair to be comparable. We selected 100 text pairs per language pair by randomly choosing 10 pairs from each decile of the retrieval tool's scoring range above the threshold (the overall aim of the study was to assess the retrieval tool as well as coder agreement).

In both the pilot and full annotation experiments the work was carried out by annotators who were members of the Accurat project team. They worked remotely using the web-based interface and the guidelines as their only source of information on the task.

2.3.2. Results

Inter-annotator agreement figures for the different categories in our scheme are shown in Table 2. Each row reports the results for two annotators' judgements over the 100 document pairs for one language. For two of the eight language pairs (Croatian and Slovenian) there were three annotators, and in these cases we report agreement results for each pair of annotators. The human judges were asked a series of questions from a set of seven questions. For each question there is a pair of columns in the table. The first column in the pair reports the percentage agreement of the two annotators on this question; the second column the raw score from which the percentage agreement was calculated, i.e. the number of document pairs for which their answer to this question was the same divided by the number of document pairs they were asked to judge. The final two rows present average percentage agreement and average Cohen's kappa scores plus the standard deviation in the kappa scores over annotator pairs.

Note that as we go across the table the denominator of the raw score goes down. This is because depending on the answers to an earlier question, a later question may not be asked. For example, if an annotator judges a document pair to be about the same news event, they will not subsequently be asked whether the document pair is about the same news event type. Furthermore, since annotators' judgement to earlier questions may diverge, only one of them may be asked a later question and in this circumstance we cannot, of course, report an agreement figure for the later questions for that document pair.

Also note that results are not given for all 100 document pairs for each language (see denominator in the "Is News Story?" raw score column). For Croatian and Romanian this is because, while all of the automatically selected document pairs had passed our language identification filters, in some cases collected documents were not actually in Croatian (frequently confused with Serbian) or Romanian. For Estonian, this is because for the time period chosen from which to gather comparable news texts, our news gathering tool simply could not find 100 text pairs that exceeded its comparability threshold.

2.3.3. Discussion

Across the seven questions asked of the human assessors, agreement ranged from 73 to 93.4 percent, the average being 81%. The two questions with the lowest percentage agreement were question 3 (75.1%), which asks whether the two stories share the same focal event, and question 6 (73%), which asks whether the two stories share the same focal event type. Both of these questions centre on the notion of focal event. Lower agreement here could result from poor annotator understanding of the notion, stemming from lack of training or careful reading of the guidelines or from lack of clarity in the guidelines, or from inherent difficulty with the notion itself. Higher agreement on this notion in the monolingual pilot where there was more discussion between the annotators and scheme developers and higher data quality control suggests that better agreement is obtainable. Highest agreement was found for question 4, which asks whether the two stories have any quotes in common. This is a relatively straightforward question to answer, so the high level of agreement is not surprising.

In order to further assess the level of annotator agreement we computed the Cohen's kappa for each annotator pair and each question. The kappa scores for each question, averaged across the language pairs, range from 0 (question 6) to .6 (questions 2 and 4). While these scores would generally be interpreted as not indicating strong agreement between annotators, there are several reasons for not attaching

[1] http://www.accurat-project.eu/. The languages the project is working with are: German (DE), Greek (EL), English (EN), Estonian (ET), Croatian (HR), Latvian (LT), Lithuanian (LV), Romanian (RO) and Slovenian (SL)

Language pair	Is News Story?		Same News Events?		Same Focal Event?		Quotes in Common?		Same News Event Type?		Same Focal Event Type?		Background in Common?	
DE-EN	81	81/100	89.7	70/78	75	36/48	91.7	44/48	86.4	19/22	100	22/22	90.9	20/22
EL-EN	88	88/100	79.5	66/83	93.5	43/46	100	46/46	80	16/20	80	16/20	60	12/20
ET-EN	88.5	69/78	88.3	53/60	83.3	25/30	96.7	29/30	82.6	19/23	78.3	18/23	95.7	22/23
HR-EN(a1-a2)	83.7	77/92	69	49/71	70	14/20	100	20/20	82.8	24/29	37.9	11/29	75.9	22/29
HR-EN(a1-a3)	90.2	83/92	89.2	66/74	65.7	23/35	94.3	33/35	67.7	21/31	93.5	29/31	77.4	24/31
HR-EN(a2-a3)	82.6	76/92	67.1	47/70	66.7	14/21	85.7	18/21	73.1	19/26	30.8	8/26	76.9	20/26
LT-EN	92	92/100	86.7	78/90	67.2	43/64	95.3	61/64	57.1	8/14	78.6	11/14	92.9	13/14
LV-EN	92	92/100	68.9	62/90	62.5	20/32	96.9	31/32	80	24/30	73.3	22/30	90	27/30
RO-EN	92.7	90/97	86.5	77/89	84.8	56/66	93.9	62/66	81.8	9/11	63.6	7/11	81.8	9/11
SL-EN(a1-a2)	76	76/100	85.3	64/75	81.1	30/37	89.2	33/37	85.2	23/27	77.8	21/27	92.6	25/27
SL-EN(a1-a3)	95	95/100	74.7	71/95	71.4	20/28	92.9	26/28	72.1	31/43	81.4	35/43	81.4	35/43
SL-EN(a2-a3)	75	75/100	69.9	51/73	80	20/25	84	21/25	73.1	19/26	80.8	21/26	84.6	22/26
Average	86.4		79.6		75.1		93.4		76.8		73.0		83.3	
Avg κ & σ	0.3	0.2	0.6	0.17	0.2	0.249	0.6	0.327	0.5	0.19	0	0.13	0.5	0.297

Table 2: % agreement for annotator responses, for different language text pairs.

too much weight to them. Kappa scores tend to be higher a) the more classes there are to assign observations to and b) the more equiprobable the class assignments are. In the current case there are just two classes per question, the minimum possible, and the classes are not at all equiprobable. In some cases, for example, nearly all the data is in one class about which the annotators may mostly agree; yet if the annotators disagree about the small number of examples outside the class kappa may be very low, despite high percentage agreement (in one case for question 1 we have 95% agreement and a kappa score of -0.02 since there is no agreement on one class). This skewed distribution across classes arises because we have used data provided by a tool which attempts to select only comparable news article pairs – a better test would involve choosing a more equal distribution across classes, though it is not clear how to do this without bias. There are also cases where the overall judgement sets are quite small – e.g. for question 6 there are an average of 25 document pairs per language pair. These small sample sizes render any statistic computed over them questionable. Finally, our discussion so far has been in terms of average kappa scores per question across all language pairs. Looking in more detail we find even more variation, with kappa scores for individual annotator pairs for single questions ranging from -.08 to 1; standard deviations of kappa scores across languages for the same question are for all questions.

We conclude that agreement is good — on average annotators will agree on 4 out of 5 judgements. To make the assessment of inter-annotator agreement more robust would require more data better distributed over the classes (i.e. data should be assembled to assess the scheme not by using a tool whose aim is to skew the data as far as possible) and more annotators. Furthermore we believe that better agreement between annotators could be obtained by adopting ideas from crowdsourcing of annotation, where annotators must demonstrate competence on a test that ensures their understanding of the task before their judgements are recorded and where difficult cases are repeatedly annotated until a reliable judgement emerges. More analysis of particular cases of disagreement and discussion with annotators will help us refine the methodology further.

3. Assessing Shared Content

In the previous section we introduced a scheme which permits news text pairs to be characterized according to how news events they report are related. Our hypothesis is that those text pairs that are most closely related, i.e. those that share focal events, will exhibit more content overlap than those that are less closely related, e.g. those that report unrelated news events. To test this hypothesis we need a method to quantify content overlap. In this section we first illustrate the challenge that finding such a method presents by discussing the sorts of phenomena we need to deal with, then discuss previous related work on approaches to assessing content overlap in text pairs, next describe the method we have adopted to determine shared content and finally discuss what we have discovered from some preliminary work on multiply annotating sample news texts according to our method.

3.1. Shared Content in Comparable Corpora

Our analysis of pairs of related news texts suggests that there are limited examples of sentence pairs which can be judged as semantically equivalent, i.e. where more or less all and only the information content expressed in one sentence is expressed by the other sentence. We do however find many examples of sentence pairs where shared content is expressed via sub-sentential units. In particular we find the following: clausal equivalence (see example (1)); phrasal equivalence, including: noun phrase-noun phrase, verb phrase-verb phrase (e.g. (2)) or noun phrase-verb phrase (e.g. (3)); and one to many sentence relationships, where the different sub-parts of one sentence, match individual phrases or clauses in multiple sentences (e.g. (4)).

(1) (a) ***The broadcast will include both men's and women's singles finals, which are taking place over the first weekend in July, and will only be shown on the BBC HD channel****, free to cable or satellite subscribers, as well as customers with Freeview HD boxes.*

(b) ***Both the men's and ladies' finals - held 2 and 3 July - will be broadcast on BBC HD, marking the 125th anniversary of the tennis tournament.***

(2) (a) ***Telephone service was down*** *in the city and throughout the area where the quake was felt*

(b) *Residents in the southwestern states of Oaxaca and Guerrero and the eastern state of Veracruz reported that* ***phone service had been knocked out****.*

(3) (a) ***A strong 7.4-magnitude earthquake*** *hit southern Mexico on Tuesday, damaging some 800 homes near the epicenter and swaying tall buildings and spreading fear and panic hundreds of miles away in the capital of Mexico City.*

(b) ***The quake had a magnitude of 7.4****, according to the U.S. Geological Survey.*

(4) (a) ***Esa modified the big antenna to widen its beam, and also reduced the power of the transmission to match the type of X-band signal Phobos-Grunt would have expected to receive nearer the Red Planet****.*

(b) (i) ***The agency had to modify its 15m dish*** *in Perth to get through to Phobos-Grunt.*

(ii) *This required* ***widening the antenna's beam*** *to catch the probe in its uncertain orbit.*

(iii) *Perth also* ***reduced the power of the transmission to make it more like the sort of faint X-band signal the craft would expect to hear at Mars****.*

Note that while there are cases of straightforward synonymy (e.g. (5)), we do not consider these to count as shared information content – as we elaborate further below we take shared content to be essentially propositional, i.e. "the same thing being said about the same thing", so that bare referential terms, even it referring to the same real world entity, do not count.

(5) (a) *Some* ***consumers*** *in online discussions have cited high temperatures with the iPad.*

(b) *If* ***customers*** *have any concerns, they should contact AppleCare*

The problem then is how to define an annotation task that will allow us to identify those sentences pairs which contain such examples of sub-sentential fragments in such a way that is feasible and likely to produce a useful evaluation resource for applications such as SMT, paraphrasing and textual entailment.

3.2. Previous Work

In their work on the creation of the Microsoft Research Paraphrase Corpus, Dolan et al (2004) developed guidelines to help annotators assess whether two sentences, from related news texts, could be considered as more or less, "semantically equivalent". Their aim was to develop a corpus rich in paraphrases, comprising sentence pairs which could be judged as more or less full paraphrases of each other. They gave the judges a very restricted selection of the overall text content, having used heuristics to select likely pairs of related sentences (the first two sentences in a text), and then further string-based filtering techniques to refine their candidate sentence pairs. While we share the aim of finding semantically equivalent sentences, their work differs from ours in two important respects. First, they were concerned with semantic equivalence at the level of full sentences, while we are concerned with it at the sub-sentential level. Second, they only considered sentences that occurred as one of the first two sentences in a text, while we consider equivalent content drawn for any position in the texts. Their emphasis can be viewed as one of precision, ours as one of recall as well as precision.

An alternative approach to identifying shared information content in texts comes from evaluation work in the summarisation community. Motivated by the desire to assess content similarity between multiple summaries rather than word-based similarity, researchers such as Nenkova, Passonneau and McKeown (2007) and Teufel and van Halteren (2004) have proposed approaches based on the idea of sub-sentential information units they call *summary content units (SCUs)* or *factoids*, respectively. In this approach human judges analyze sentences drawn from multiple summaries and identify the common SCUs or factoids – informational chunks that in context are deemed to be shared and which may be realized by an entire sentence or a single word. The common information units are given a label and a natural language gloss by the annotator, who also notes the textual extent expressing the information in the two sentences. They go on to show how SCUs or factoids emerging from multiple reference summaries can be used in summary evaluation by assessing how many of the information units shared by the most reference summaries are found in the summary under evaluation. While this aspect of their work is not of relevance to us, since our task differs, key features of their approach from our perspective are: (1) it works at the level of meaning rather than surface similarity, which is essential for dealing with comparable texts; (2) it works below the level of the sentence, which is important since, as we have noted, much shared content in comparable news texts is expressed via sub-sentential text fragments; (3) it does not require analysis into an abstract formal meaning representation or to an agreed level of semantic primitives, but rather just for annotators to recognize there is common information and to be able to mark its textual extent; (4) extensive empirical work has shown judgements based on this approach to be stable and the Pyramid method, which is based upon it, has been widely adopted as an evaluation measure in summarization research.

Given these strengths we decided to adapt this method for identifying shared information units to our task of determining shared content in comparable news texts.

3.3. Annotating Shared Content

We have developed an annotation task for identifying sentence pairs in two comparable news texts which contain shared semantic content based on the notion of shared units of information, which we will refer to as "SUIs". These UIs

Text Id	Sentence Pairs	Shared UI Agree=y	Shared UI Agree=no	Shared UI Disagree	Full Sentence Equivalence
1	64	14/21.9%	49/76.6%	1/1.6%	1/1.6%
2	99	13/13.1%	85/85.9%	1/1.0%	0/0.0%
TOTAL	163	27/16.6%	134/82.2%	2/1.2%	1/0.6%

Table 3: Two Annotator Agreement on UI Annotation over Two Sample Text Pairs

are like the SCU's or factoids discussed above, which are never precisely defined, but we make explicit the requirement that a UI be propositional, i.e. "say something about something". This does not mean UIs have to take an obvious NP VP form of expression, but there should be some predication involved, not merely reference (so *The table* is not an UI while *The table is broken* or *The broken table* is). Having split paired texts into their constituent sentences and computed all possible sentence pairings, we ask annotators to examine each sentence pair and to give a binary judgement as to whether the sentences contain shared SUIs. As in the Pyramid method we do not ask annotators to enumerate all candidate UI's in each sentence by reference to a formal definition and then determine their overlap. Rather, we advise that for sentences that appear to be saying similar things, annotators first identify potential UIs by decomposing likely chunks of overlapping information into more elementary propositions, e.g. by identifying who did what to whom, with what and where, etc., and then determine which of these is shared. We also ask that annotators record a brief description in natural language of any UI's, which are common to both sentences (this helps to focus the task and encourages active sentence analysis). In addition we ask participants to mark, in each sentence, the textual extent that supports the UI. At this stage, we insist on a contiguous extent, so we expect additional text which does not contribute to SUIs to be included in the results.

(6) (a) *Two men* ***were rescued by an RAF helicopter co-piloted by Prince William***.

(b) ***Prince William, who is a search and rescue helicopter co-pilot at RAF Valley, took part in the rescue***.

For example, in (6) we identify a SUI, which can be glossed as *Prince William, who is a helicopter co-pilot, took part in a rescue operation*. In the text we have bolded the text that supports the SUIs in each case.

Note that the presence of SUIs does not imply full semantic equivalence. In cases of one way textual entailment we would expect a SUI to be annotated, where the SUI is the entailed proposition. For example, the two highlighted expressions in (7) both entail that a huge wave has hit a cargo boat, but in (b) we see more specific details about the blow, i.e. that it snapped the hull.

(7) (a) ***The 81-metre cargo carrier sank 10 miles west of the Lleyn peninsula in north Wales after being hit by an "enormous wave"***.

(b) ***Five crew from a cargo ship are feared dead after a huge wave snapped the vessel's hull*** *in stormy seas off the coast of north Wales.*

Guidelines for the task provide further details on how to handle, e.g., anaphora, minor variations in quantities and measures, attribution, complex entailment, extent selection for supporting UIs, and so on.

3.4. Pilot Annotation

To form an initial impression of the difficulty and robustness of the annotation task, two annotators analyzed all sentence pairings for two news text pairs on the same focal event.

Our analysis of the results of this exercise (see Table 3) showed that the task of SUI recognition for sentence pairs can be carried out with a high rate of agreement between annotators, with disagreement recorded for just 2 of the total 163 pairs examined, i.e. there was 98.8% agreement on which sentence pairs contained SUIs. In 27 sentence pairs both annotators agreed the was an underlying SUI (nearly 17% of the total), a significant improvement on the number of sentences which both annotators judged to have full semantic equivalence (i.e., 1 or 0.6%).

In the 27 sentence pairs where annotators had agreed on the presence of a SUI, we found good agreement on the position of supporting text extents. There were differences in only 6 of the 54 extent comparisons, and these occurred in only 4 of the 27 SUI sentence pairs, (giving us a total of 23 agreed pairs of selected text extents). Differences ranged from 2-6 words long, with one outlier of a 20 word difference. We note that supporting extents (with full agreement) ranged from a minimum of 2 to a maximum of 28 words.

We examined the 23 agreed pairs of sub-sentential text extents, and found the set to include many paraphrases. Among these we found examples of various types of paraphrase alternation (similar to those reported by Dolan et al. (2004)): anaphora, word and phrase re-ordering, including active/passive inversion; textual entailment and elaboration. These results are promising, but of course a much larger sample of text pairs needs to be annotated before any solid conclusions can be drawn.

4. Conclusion

We have presented initial stages of work on a multi-stage programme for assessing the comparability of news texts with a view to gaining insights that will enable us to better exploit them for language processing tasks, such as machine translation. First, based on the observation that the notion of being "on the same news event" that underlies much current work on creating comparable corpora of news texts is vague and that text pairs meeting this criterion exhibit a huge variation in content overlap, we proposed a scheme for classifying news text pairs according to the degree of relatedness of the events they report. We investi-

gated how robust this scheme is through a multi-lingual annotation exercise involving comparable news texts for eight language pairs. This exercise showed reasonable agreement amongst annotators, but more work is necessary to ensure annotators have understood the task, to improve data selected for agreement and in analysis of divergences.

Second, we proposed a method for identifying shared content in news text pairs which is an adaptation of methods developed in summarization evaluation, such as the Pyramid method (Nenkova et al., 2007), which also address the problem of identifying shared content amongst multiple texts where the surface expression may differ. Our method is based on the idea of "shared units of information" between two sentences which annotators identify and then mark supporting evidence for in the respective sentences. We described our method and then presented a preliminary exercise in which two annotators followed the method to annotate a small number of example text pairs. Results are very encouraging both in terms of the amount of shared information found at the sub-sentential level, and consequently text strings that may be considered to be paraphrases, and in terms of annotator agreement.

Of course much more extensive exercises are necessary to properly validate the approach and to establish in detail the character of the resulting sub-sentential aligned fragments, and their utility for various language processing tasks, such as SMT, paraphrase acquisition and textual entailment. As part of this work, we plan to to determine how alignable at the word level the text extents marked in our shared content annotation method are, using a method such as the one Cohn et al. (2008) proposed for aligning the full sentence paraphrases present in the Microsoft paraphrase corpus (Dolan et al., 2004). This will give insights into how much value the word strings expressing shared content in comparable news texts could be to current SMT systems.

The third stage in our programme is to assess how the amount of shared content between news text pairs varies with the categories of our news story event relatedness scheme. The underlying hypothesis here is that news texts that share the same focal event will contain more shared content than those on the same news event, but whose focal events differ (if true this would mean efforts should concentrate on collecting news stories with the same focal event in order to maximize the amount of shared content in comparable news corpora). To do this we must annotate for shared content enough text pairs in different categories of the news event type relatedness scheme to enable us to confirm or reject the hypothesis about correlation of quantity of shared content and categories of the scheme.

Finally, if results are positive in the second and third stages, the challenge is to create algorithms to gather same focal event news text pairs automatically and to identify and align the phrases expressing shared content within them.

Acknowledgments

We would like to acknowledge funding from the European Commission FP7 programme within the ACCURAT project (`http://www.accurat-project.eu/`) and thank our project partners for their help with the annotation exercise reported in Section 2.3.

5. References

Ahmet Aker, Evangelos Kanoulas, and Robert Gaizauskas. 2012. A light way to collect comparable corpora from the web. In *Proc. LREC 2012*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proc. NAACL-HLT 2003*, pp. 16–23.

Allan Bell. 1998. The discourse structure of news stories. In Allan Bell and Peter Garrett, editors, *Approaches to Media Discourse*. Blackwell Publishers.

Martin Braschler and Peter Schäuble. 1998. Multilingual information retrieval based on document alignment techniques. In *ECDL*, pp. 183–197.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for development and evaluation of paraphrase systems. *Computational Lingustics*, 34(4):597–614.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proc. 20th Int. Conference on Computational Linguistics*.

Jonathan G. Fiscus and George R. Doddington. 2002. Topic detection and tracking evaluation overview. In James Allan, ed., *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Parallel Text Processing*, pp. 1–17. Springer.

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience. *Journal of Narrative & Life History*, 7(1-4):3–38.

Bo Li, Éric Gaussier, and Akiko N. Aizawa. 2011. Clustering comparable corpora for bilingual lexicon extraction. In *ACL 2011 (Short Papers)*, pp. 473–478.

Dragos S. Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Compututational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proc. COLING/ACL 2006*, pp. 81–88.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4, May.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proc. of the 2nd lnt. Conf. on Human Language Technology*, HLT '02, pp. 313–318.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proc. EMNLP 2004*, pp. 419–426.

T.A. van Dijk. 1985. Structures of news in the press. In T.A. van Dijk, editor, *Discourse and Communication*, pp. 69–93. De Gruyter, Berlin.