Multi-word expressions in textual inference: Much ado about nothing?

Marie-Catherine de Marneffe
Linguistics Department
Stanford University
Stanford, CA
mcdm@stanford.edu

Sebastian Padó
Institut für Maschinelle
Sprachverarbeitung
Stuttgart University, Germany
pado@ims.uni-stuttgart.de

Christopher D. Manning
Computer Science Department
Stanford University
Stanford, CA
manning@stanford.edu

Abstract

Multi-word expressions (MWE) have seen much attention from the NLP community. In this paper, we investigate their impact on the recognition of textual entailment (RTE). Using the manual Microsoft Research annotations, we first manually count and classify MWEs in RTE data. We find few, most of which are arguably unlikely to cause processing problems. We then consider the impact of MWEs on a current RTE system. We are unable to confirm that entailment recognition suffers from wrongly aligned MWEs. In addition, MWE alignment is difficult to improve, since MWEs are poorly represented in state-of-the-art paraphrase resources, the only available sources for multi-word similarities. We conclude that RTE should concentrate on other phenomena impacting entailment, and that paraphrase knowledge is best understood as capturing general lexico-syntactic variation.

1 Introduction

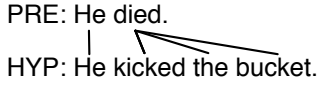
Multi-word expressions (MWEs) can be defined as "idiosyncratic interpretations that cross word boundaries", such as *traffic light* or *kick the bucket*. Called a "pain in the neck for NLP", they have received considerable attention in recent years and it has been suggested that proper treatment could make a significant difference in various NLP tasks (Sag et al., 2002). The importance attributed to them is also reflected in a number of workshops (Bond et al., 2003; Tanaka et al., 2004; Moirón et al., 2006; Grégoire et al., 2007). However, there are few detailed breakdowns of the benefits that improved MWE handling provides to applications.

This paper investigates the impact of MWEs on the "recognition of textual entailment" (RTE) task (Dagan et al., 2006). Our analysis ties in with the pivotal question of what types of knowledge are beneficial for RTE. A number of papers have suggested that *paraphrase knowledge* plays a very important role (Bar-Haim et al., 2005; Marsi et al., 2007; Dinu and Wang, 2009). For example, Bar-Haim et al. (2005) conclude: "Our analysis also shows that paraphrases stand out as a dominant contributor to the entailment task."

The term "paraphrase" is however often construed broadly. In Bar-Haim et al. (2005), it refers to the ability of relating *lexico-syntactic* reformulations such as diathesis alternations, passivizations, or symmetrical predicates (*X lent his BMW to Y*/*Y borrowed X's BMW*). If "paraphrase" simply refers to the use of a language's lexical and syntactic possibilities to express equivalent meaning in different ways, then paraphrases are certainly important to RTE. But such a claim means little more than that RTE can profit from good understanding of syntax and semantics. However, given the abovementioned interest in MWEs, there is another possibility: does success in RTE involve proper handling of MWEs, such as knowing that *take a pass on* is equivalent to *aren't purchasing*, or *kicked the bucket* to *died*? This seems not too far-fetched: Knowledge about MWEs is under-represented in existing semantic resources like WordNet or distributional thesauri, but should be present in paraphrase resources, which provide similarity judgments between phrase pairs, including MWEs.

The goal of our study is to investigate the merits of this second, more precise, hypothesis, measuring the impact of MWE processing on RTE. In the absence of a universally accepted definition of MWEs, we define MWEs in the RTE setting as *multi-word alignments*, i.e., words that participate in more than one word alignment link between premise and hypothesis:

(1)



The exclusion of MWEs that do not lead to multi-word alignments (i.e., which can be aligned word by word) is not a significant loss, since these cases are unlikely to cause significant problems for RTE. In addition, an alignment-based approach has the advantage of generality: Almost all existing RTE models *align* the linguistic material of the premise

and hypothesis and base at least part of their decision on properties of this alignment (Burchardt et al., 2007; Hickl and Bensley, 2007; Iftene and Balahur-Dobrescu, 2007; Zanzotto et al., 2007).

We proceed in three steps. First, we analyze the Microsoft Research (MSR) manual word alignments (Brockett, 2007) for the RTE2 dataset (Bar-Haim et al., 2006), shedding light on the relationship between alignments and multi-word expressions. We provide frequency estimates and a coarse-grained classification scheme for multi-word expressions on textual entailment data. Next, we analyze two widely used types of paraphrase resources with respect to their modeling of MWEs. Finally, we investigate the impact of MWEs and their handling on practical entailment recognition.

2 Multi-Word Expressions in Alignment

Almost all textual entailment recognition models incorporate an alignment procedure that establishes correspondences between the premise and the hypothesis. The computation of word alignments is usually phrased as an optimization task. The search space is based on lexical similarities, but usually extended with *structural biases* in order to obtain alignments with desirable properties, such as the contiguous alignment of adjacent words, or the mapping of different source words on to different target words. One prominent constraint of the IBM word alignment models (Brown et al., 1993) is *functional alignment*, that is each target word is mapped onto *at most* one source word. Other models produce only *one-to-one alignments*, where both alignment directions must be functional.

MWEs that involve many-to-many or one-to-many alignments like Ex. (1) present a problem for such constrained word alignment models. A functional alignment model can still handle cases like Ex. (1) correctly in one direction (from bottom to top), but not in the other one. One-to-one alignments manage neither. Various workarounds have been proposed in the MT literature, such as computing word alignments in both directions and forming the union or intersection. Even if an alignment is technically within the search space, accurate knowledge about plausible phrasal matches is necessary for it to be assigned a high score and thus identified.

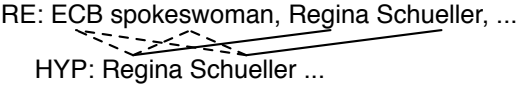
3 MWEs in the RTE2 Dataset

In the first part of our study, we estimate the extent to which the inability of aligners to model one-to-many and many-to-many correspondences is an issue. To do so, we use the Microsoft Research manual alignments for the RTE2 data. To date, the MSR data constitutes the only gold standard alignment corpus publicly available. Since annotators were not constrained to use one-to-one alignments, we assume that the MSR alignments contain multi-word alignments where appropriate.

		CARDINALITY	
		M-to-M	1-to-M
DECOMPOSABLE?	yes	(1)	(3)
	no	(2)	(4)
OTHER			(5), (6), (7)

Table 1: MWEs categories and definition criteria (M-to-M: many-to-many; 1-to-M: one-to-many).

From the MSR data, we extract all multi-word alignments that fall outside the scope of "functional" alignments, i.e., alignments of the form "many-to-many" or "one-to-many" (in the direction hypothesis-premise). We annotate them according to the categories defined below. The MSR data distinguishes between SURE and POSSIBLE alignments. We only take the SURE alignments into account. While this might mean missing some multi-word alignments, we found many "possible" links to be motivated by the desire to obtain a high-coverage alignment, as Ex. 2 shows:

(2)


Here, the hypothesis words "Regina Schueller" are individually "sure"-aligned to the premise words "Regina Schueller" (solid lines), but are also both "possible"-linked to "ECB spokeswoman" (dashed lines). This "possible" alignment can be motivated on syntactic or referential grounds, but does not indicate a correspondence in meaning (as opposed to reference).

3.1 Analysis of Multi-Word Expressions

Table 1 shows the seven categories we define to distinguish the different types of multi-word alignments. We use two main complementary criteria for our annotation. The first one is the cardinality of the alignment: does it involve phrases proper on both sides (many-to-many), or just on one side (one-to-many)? The second one is decomposability: is it possible to create one or more one-to-one alignments that capture the main semantic contribution of the multi-word alignment? Our motivation

for introducing this criterion is that even aligners that are unable to recover the complete MWE have a chance to identify the links crucial for entailment if the MWE is decomposable (categories (1) and (3)). This is not possible for the more difficult non-decomposable categories (2) and (4). The remaining categories, (5) to (7), involve auxiliaries, multiple mentions, and named entities, which are not MWEs in the narrow sense. We will henceforth use the term "true MWEs" to refer to categories (1)–(4), as opposed to (5)–(7).

The criteria we use for MWE categorization are different from the ones adopted by Sag et al. (2002). Sag et al.'s goal is to classify constructions by their range of admissible variation, and thus relies heavily on syntactic variability. Since we are more interested in semantic properties, we base our classes on alignment patterns, complemented by semantic decomposability judgments (which reflect the severity of treating MWEs like compositional phrases). As mentioned in Section 1, our method misses MWEs aligned with one-to-one links; however, the use of a one-to-one link by the annotation can be seen as evidence for decomposability.

A. Multiple words on both sides

(1) **Compositional phrases (CP)**:
Each word in the left phrase can be aligned to one word in the right phrase, e.g., *capital punishment* → *death penalty* for which *capital* can be aligned to *death* and *punishment* to *penalty*.

(2) **Non-compositional phrases (NCP)**:
There is no simple way to align words between the two phrases, such as in *poorly represented* → *very few* or *illegally entered* → *broke into*.

B. One word to multiple words

(3) **Headed multi-word expressions (MWEH)**:
A single word can be aligned with one token of an MWE: e.g., *vote* → *cast ballots* where *ballots* carries enough of the semantics of *vote*.

(4) **Non-headed MWEs (MWENH)**:
The MWE as a whole is necessary to capture the meaning of the single word, which doesn't align well to any individual word of the MWE: e.g., *ferry* → *passenger vessel*.

(5) **Multiple mentions (MENTION)**:
These alignments link one word to multiple occurrences of the same or related word(s) in the text, e.g., *military* → *forces ... Marines*, *antibiotics* → *antibiotics ... drug*.

(6) **Parts of named entities (PART)**:
Each element of a named entity is aligned to the whole named entity: e.g., *Shukla* → *Nidhi Shukla*. This includes the use of acronyms or abbreviations on one side and their spelled-out forms on the other side, such as *U.S.* → *United States.*

(7) **Auxiliaries (AUX)**:
The last category involves the presence of an auxiliary: e.g., *were* → *are being*.

Status	Category	RTE2 dev	RTE2 test
decomp.	CP	5	0
	MWEH	40	31
non-decomp.	NCP	6	0
	MWENH	30	29
Subtotal: True MWEs		**81**	**60**
other	MENTION	26	48
	PART	82	54
	AUX	0	2
Total: All MWEs		189	164

Table 2: Frequencies of sentences with different multi-word alignment categories in MSR data.

Initially, one of the authors used these categories to analyze the complete RTE2 MSR data (dev and test sets). The most difficult distinction to draw was, not surprisingly, the decision between decomposable multi-word alignments (categories (1) and (3)) and non-decomposable ones (categories (2) and (4)). To ascertain that a reliable distinction can be made, another author did an independent second analysis of the instances from categories (1) through (4). We found moderate inter-annotator agreement (κ = 0.60), indicating that not all, but most annotation decisions are uncontroversial.

3.2 Distribution of Multi-Word Expressions

Table 2 shows the distribution in the MSR data of all alignment categories. Our evaluation will concentrate on the "true MWE" categories (1) to (4): CP, NCP, MWEH and MWENH.[1]

[1]The OTHER categories (5) to (7) can generally be dealt with during pre- or post-processing: Auxiliary-verb combinations (cat. 7) are usually "headed" so that it is sufficient to align the main verb; multiple occurrences of words referring to the same entity (cat. 5) is an anaphor resolution problem; and named-entity matches (cat. 6) are best solved by using a named entity recognizer to collapse NEs into a single token.

In RTE2 dev and test, we find only 81 and 60 true MWEs, respectively. Out of the 1600 sentence pairs in the two datasets, 8.2% involve true MWEs (73 in RTE2 dev and 58 in RTE2 test). On the level of word alignments, the ratio is even smaller: only 1.2% of all SURE alignments involve true MWEs. Furthermore, more than half of them are decomposable (MWEH/CP). Some examples from this category are ("heads" marked in boldface):

sue→ *file* ***lawsuits*** *against*
diseases → *liver* ***cancer***
Barbie → ***Barbie*** *doll*
got → *was* ***awarded*** *with*
works → *executive* ***director***
military → *naval* ***forces***

In particular when light verbs are involved (*file lawsuits*) or when modification adds just minor meaning aspects (*executive director*), we argue that it is sufficient to align the left-hand expression to the "head" in order to decide entailment.

Consider, in contrast, these examples from the non-decomposable categories (MWENH/NCP):

politician → *presidential candidate*
killed → *lost their lives*
shipwreck → *sunken ship*
ever → *in its history*
widow → *late husband*
sexes → *men and women*

These cases span a broad range of linguistic relations from pure associations (*widow/late husband*) to collective expressions (*sexes/men and women*). Arguably, in these cases aligning the left-hand word to any single word on the right can seriously throw off an entailment recognition system. However, they are fairly rare, occurring only in 65 out of 1600 sentences.

3.3 Conclusions from the MSR Analysis

Our analysis has found that 8% of the sentences in the MSR dataset involve true MWEs. At the word level, the fraction of true MWEs of all SURE alignment links is just over 1%.

Of course, if errors in the alignment of these MWEs had a high probability to lead to entailment recognition errors, MWEs would still constitute a major factor in determining entailment. However, we have argued that about half of the true MWEs are *decomposable*, that is, the part of the alignment that is crucial for entailment can be recovered with a one-to-one alignment link that can be identified even by very limited alignment models.

This leaves considerably less than 1% of all word alignments (or ∼4% of sentence pairs) where imperfect MWE alignments *are able at all* to exert a negative influence on entailment. However, this is just an upper bound – their impact is by no means guaranteed. Thus, our conclusion from the annotation study is that we do not expect MWEs to play a large role in actual entailment recognition.

4 MWEs in Paraphrase Resources

Before we come to actual experiments on the automatic recognition of MWEs in a practical RTE system, we need to consider the prerequisites for this task. As mentioned in Section 2, if an RTE system is to establish multi-word alignments, it requires a knowledge source that provides accurate semantic similarity judgments for "many-to-many" alignments (*capital punishment – death penalty*) as well as for "one-to-many" alignments (*vote – cast ballots*). Such similarities are not present in standard lexical resources like WordNet or Dekang Lin's thesaurus (Lin, 1998).

The best class of candidate resources to provide wide-coverage of multi-word similarities seems to be *paraphrase* resources. In this section, we examine to what extent two of the most widely used paraphrase resource types provide supporting evidence for the true MWEs in the MSR data. We deliberately use corpus-derived, noisy resources, since we are interested in the real-world (rather than idealized) prospects for accurate MWE alignment.

Dependency-based paraphrases. Lin and Pantel (2002)'s DIRT model collects lexicalized dependency paths with two slots at either end. Paths with similar distributions over slot fillers count as paraphrases, with the quality measured by a mutual information-based similarity over the slot fillers. The outcome of their study is the DIRT database which lists paraphrases for around 230,000 dependency paths, extracted from about 1 GB of miscellaneous newswire text. We converted the DIRT paraphrases[2] into a resource of semantic similarities between raw text phrases. We used a heuristic mapping from dependency relations to word order, and obtained similarity ratings by rescaling the DIRT paraphrase ratings, which are based on a mutual information-based measure of filler similarity, onto the range [0,1].

[2]We thank Patrick Pantel for granting us access to DIRT.

Parallel corpora-based paraphrases. An alternative approach to paraphrase acquisition was proposed by Bannard and Callison-Burch (2005). It exploits the variance inherent in translation to extract paraphrases from bilingual parallel corpora. Concretely, it observes translational relationships between a source and a target language and pairs up source language phrases with other source language phrases that translate into the same target language phrases. We applied this method to the large Chinese-English GALE MT evaluation P3/P3.5 corpus (∼2 GB text per language, mostly newswire). The large number of translations makes it impractical to store all observed paraphrases. We therefore filtered the list of paraphrases against the raw text of the RTE corpora, acquiring the 10 best paraphrases for around 100,000 two- and three-word phrases. The MLE conditional probabilities were scaled onto [0,1] for each target.

Analysis. We checked the two resources for the presence of the true MWEs identified in the MSR data. We found that overall 34% of the MWEs appear in these resources, with more decomposable MWEs (MWEH/CP) than non-decomposable ones (MWENH/NCP) (42.1% vs. 24.6%). However, we find that almost all of the MWEs that are covered by the paraphrase resources are assigned very low scores, while erroneous paraphrases (expressions with clearly different meanings) have higher scores. This is illustrated in Table 3 for the case of *poorly represented*, which is aligned to *very few* in one RTE2 sentence. This paraphrase is on the list, but with a lower similarity than unsuitable paraphrases such as *representatives* or *good*. This problem is widespread. Other examples of low-scoring paraphrases are: *another step* → *measures*, *quarantine* → *in isolation*, *punitive measures* → *sanctions*, *held a position* → *served as*, or *inability* → *could not*.

The noise in the rankings means that any alignment algorithm faces a dilemma: either it uses a high threshold and misses valid MWE alignments, or it lowers its threshold and risks constructing incorrect alignments.

poorly represented	
represented	0.42
poorly	0.07
rarely	0.06
good	0.05
representatives	0.04
very few	0.04
well	0.02
representative	0.01

Table 3: Paraphrases of "poorly represented" with scores (semantic similarities).

5 Impact of MWEs on Practical Entailment Recognition

This section provides the final step in our study: an evaluation of the impact of MWEs on entailment recognition in a current RTE system, and of the benefits of explicit MWE alignment. While the results of this experiment are not guaranteed to transfer to other RTE system architectures, or to future, improved paraphrase resources, it provides a current snapshot of the practical impact of MWE handling.

5.1 The Stanford RTE System

We base our experiments on the Stanford RTE system which uses a staged architecture (MacCartney et al., 2006). After the linguistic analysis which produces dependency graphs for premise and hypothesis, the alignment stage creates links between the nodes of the two dependency trees. In the inference stage, the system produces roughly 70 features for the aligned premise-hypothesis pair, almost all of which are implementations of "small linguistic theories" whose activation indicates lexical, syntactic and semantic matches and mismatches of different types. The entailment decision is computed using a logistic regression on these features.

The Stanford system supports the use of different aligners without touching the rest of the pipeline. We compare two aligners: a one-to-one aligner, which cannot construct MWE alignments (UNIQ), and a many-to-many aligner (MANLI) (MacCartney et al., 2008), which can. Both aligners use around 10 large-coverage lexical resources of semantic similarities, both manually compiled resources (such as WordNet and NomBank) and automatically induced resources (such as Dekang Lin's distributional thesaurus or InfoMap).

UNIQ: A one-to-one aligner. UNIQ constructs an alignment between dependency graphs as the highest-scoring mapping from each word in the hypothesis to one word in the premise, or to null. Mappings are scored by summing the alignment scores of all individual word pairs (provided by the lexical resources), plus edge alignment scores that

use the syntactic structure of premise and hypothesis to introduce a bias for syntactic parallelism. The large number of possible alignments (exponential in the number of hypothesis words) makes exhaustive search intractable. Instead, UNIQ uses a stochastic search based on Gibbs sampling, a well-known Markov Chain Monte Carlo technique (see de Marneffe et al. (2007) for details).

Since it does not support many-to-many alignments, the UNIQ aligner cannot make use of the multi-word information present in the paraphrase resources. To be able to capture some common MWEs, the Stanford RTE system was originally designed with a facility to concatenate MWEs present in WordNet into a single token (mostly particle verbs and collocations, e.g., *treat_as* or *foreign_minister*). However, we discovered that WordNet collapsing always has a negative effect. Inspection of the constructed alignments suggests that the lexical resources that inform the alignment process do not provide scores for most collapsed tokens (such as *wait_for*), and precision suffers.

MANLI: A phrase-to-phrase aligner. MANLI aims at finding an optimal alignment between phrases, defined as contiguous spans of one or multiple words. MANLI characterizes alignments as *edit scripts*, sets of edits (substitutions, deletions, and insertions) over phrases. The quality of an edit script is the sum of the quality of the individual edit steps. Individual edits are scored using a feature-based scoring function that takes edit type and size into consideration.[3] The score for substitution edits also includes a lexical similarity score similar to UNIQ, plus potential knowledge about the semantic relatedness of multi-word phrases not expressible in UNIQ. Substitution edits also use contextual features, including a distortion score and a matching-neighbors feature.[4] Due to the dependence between alignment and segmentation decisions, MANLI uses a simulated annealing strategy to traverse the resulting large search space.

Even though MANLI is our current best candidate at recovering MWE alignments, it currently has an important architectural limitation: it works on textual phrases rather than dependency tree fragments, and therefore misses all MWEs that are not contiguous (e.g., due to inserted articles or adverbials). This accounts for roughly 9% of the MWEs in RTE2 data. Other work on RTE has targeted specifically this observation and has described paraphrases on a dependency level (Marsi et al., 2007; Dinu and Wang, 2009).

[3] Positive weights for all operation types ensure that MANLI prefers small over large edits where appropriate.

[4] An adaptation of the averaged perceptron algorithm (Collins, 2002) is used to tune the model parameters.

		micro-avg		
		P	R	F_1
UNIQ	w/o para	80.4	80.8	80.6
MANLI	w/o para	77.0	85.5	81.0
	w/ para	76.7	85.4	80.8

Table 4: Evaluation of aligners and resources against the manual MSR RTE2 test annotations.

Setup. To set the parameters of the two models (i.e., the weights for different lexical resources for UNIQ, and the weights for the edit operation for MANLI), we use the RTE2 development data. Testing takes place on the RTE2 test and RTE4 datasets. For MANLI, we performed this procedure twice, with the paraphrase resources described in Section 4 once deactivated and once activated. We evaluated the output of the Stanford RTE system both on the word alignment level, and on the entailment decision level.

5.2 Evaluation of Alignment Accuracy

The results for evaluating the MANLI and UNIQ alignments against the manual alignment links in the MSR RTE2 test set are given in Table 4. We present micro-averaged numbers, where each alignment link counts equally (i.e., longer problems have a larger impact). The overall difference is not large, but MANLI produces a slightly better alignment.

The ability of MANLI to construct many-to-many alignments is reflected in a different position on the precision/recall curve: the MANLI aligner is less precise than UNIQ, but has a higher recall. Examples for UNIQ and MANLI alignments are shown in Figures 1 and 2. A comparison of the alignments shows the pattern to be expected from Table 4: MANLI has a higher recall, but contains occasional questionable links, such as *at President* → *President* in Figure 1.

However, the many-to-many alignments that MANLI produces do not correspond well to the MWE alignments. The overall impact of the paraphrase resources is very small, and their addition actually hurts MANLI's performance slightly. A more detailed analysis revealed two contrary trends. On the one hand, the paraphrase resources provide

Aligner	w/o para	w/ para
UNIQ	63.8	–
MANLI	60.6	60.6

Table 5: Entailment recognition accuracy of the Stanford system on RTE2 test (two-way task).

Aligner	w/o para	w/ para	TAC system
UNIQ	63.3	–	61.4
MANLI	59.0	57.9	57.0

Table 6: Entailment recognition accuracy of the Stanford system on RTE4 (two-way task).

beneficial information, maybe surprisingly, in the form of broad distributional similarities for *single* words that were not available from the standard lexical resources (e.g., the alignment "the company's *letter*" → "the company's *certificate*").

On the other hand, MANLI captures not one of the true MWEs identified in the MSR data. It only finds two many-to-many alignments which belong to the CP category: *aimed criticism* → *has criticised*, *European currency* → *euro currency*. We see this as the practical consequences of our observation from Section 4: The scores in current paraphrase resources are too noisy to support accurate MWE recognition (cf. Table 3).

5.3 Evaluation of Entailment Recognition

We finally evaluated the performance of the Stanford system using UNIQ and MANLI alignments on the entailment task. We consider two datasets: RTE2 test, the alignment evaluation dataset, and the most recent RTE4 dataset, where current numbers for the Stanford system are available from last year's Text Analysis Conference (TAC).

A reasonable conjecture would be that better alignments translate into better entailment recognition. However, as the results in Tables 5 and 6 show, this is not the case. Overall, UNIQ outperforms MANLI by several percent accuracy despite MANLI's better alignments. This "baseline" difference should not be overinterpreted, since it may be setup-specific: the features computed in the inference stage of the Stanford system were developed mainly with the UNIQ aligner in mind. A more significant result is that the integration of paraphrase knowledge in MANLI has no effect on RTE2 test, and even decreases performance on RTE4.

The general picture that we observe is that there is only a loose coupling between alignments and the entailment decision: individual alignments seldom matter. This is shown, for example, by the alignments in Figures 1 and 2. Even though MANLI provides a better overall alignment, UNIQ's alignment is "good enough" for entailment purposes. In Figure 1, the two words UNIQ leaves unaligned are a preposition (*at*) and a light verb (*aimed*), both of which are not critical to determine whether or not the premise entails the hypothesis.

This interpretation is supported by another analysis, where we tested whether entailments involving at least one true MWE are more difficult to recognize. We computed the entailment accuracy for all applicable RTE2 test pairs (7%, 58 sentences). The accuracy on this subset is 62% for the MANLI model without paraphrases, 64% for the MANLI model with paraphrases, and 74% for UNIQ. The differences from the numbers in Table 5 are not significant due to the small size of the MWE sample, but we observe that the accuracy on the MWE subset tends to be *higher* than on the whole set (rather than lower). Futhermore, even though we finally see a small beneficial effect of paraphrases on the MANLI aligner, the UNIQ aligner, which completely ignores MWEs, still performs substantially better.

Our conclusion is that wrong entailment decisions rarely hinge on wrongly aligned MWEs, at least with a probabilistic architecture like the Stanford system. Consequently, it suffices to recover the most crucial alignment links to predict entailment, and the benefits associated with the use of a more *restricted* alignment formulation, like the one-to-one alignment formulation of UNIQ, outweighs those of more powerful alignment models, like MANLI's phrasal alignments.

6 Conclusions

We have investigated the influence of multi-word expressions on the task of recognizing textual entailment. In contrast to the widely held view that proper treatment of MWEs could bring about a substantial improvement in NLP tasks, we found that the importance of MWEs in RTE is rather small. Among the MWEs that we identified in the alignments, more than half can be captured by one-to-one alignments, and should not pose problems for entailment recognition.

Furthermore, we found that the remaining MWEs are rather difficult to model faithfully. The MSR MWEs are poorly represented in state-of-the-

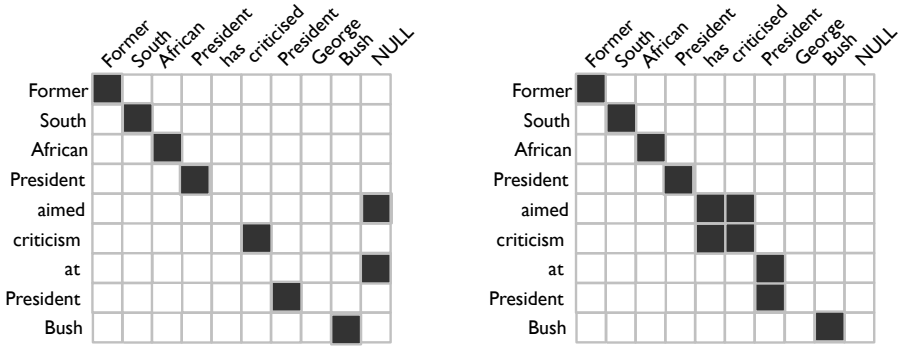


Figure 1: UNIQ (left) and MANLI (right) alignments for problem 483 in RTE2 test. The rows represent the hypothesis words, and the columns the premise words.

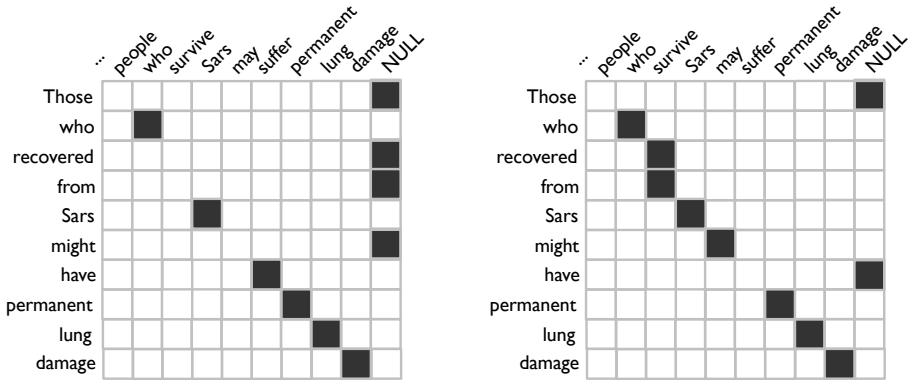


Figure 2: UNIQ (left) and MANLI (right) alignments for problem 1 in RTE2 test.

art lexical resources, and when they are present, scoring issues arise. Consequently, at least in the Stanford system, the integration of paraphrase knowledge to enable MWE recognition has made almost no difference either in terms of alignment accuracy nor in entailment accuracy. Furthermore, it is not the case that entailment recognition accuracy is worse for sentences with "true" MWEs. In sum, we find that even though capturing and representing MWEs is an interesting problem in itself, MWEs do not seem to be such a pain in the neck – at least not for textual entailment.

Our results may seem to contradict the results of many previous RTE studies such as (Bar-Haim et al., 2005) which found paraphrases to make an important contribution. However, the beneficial effect of paraphrases found in these studies refers not to an alignment task, but to the ability of relating *lexico-syntactic* reformulations such as diathesis alternations or symmetrical predicates (*buy/sell*). In the Stanford system, this kind of knowledge is already present in the features of the inference stage. Our results should therefore rather be seen as a clarification of the complementary nature of the paraphrase and MWE issues.

In our opinion, there is much more potential for improvement from better estimates of semantic similarity. This is true for phrasal similarity, as our negative results for multi-word paraphrases show, but also on the single-word level. The 2% gain in accuracy for the Stanford system here over the reported TAC RTE4 results stems merely from efforts to clean up and rescale the lexical resources used by the system, and outweighs the effect of MWEs. One possible direction of research is conditioning semantic similarity on *context*: Most current lexical resources characterize similarity at the lemma level, but true similarities of word or phrase pairs are strongly context-dependent: *obtain* and *be awarded* are much better matches in the context of *a degree* than in the context of *data*.

Acknowledgments

We thank Bill MacCartney for his help with the MANLI aligner, and Michel Galley for the parallel corpus-based paraphrase resource. This paper is based on work funded in part by DARPA through IBM. The content does not necessarily reflect the views of the U.S. Government, and no official endorsement should be inferred.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 597–604, Ann Arbor, MI.

Roy Bar-Haim, Idan Szpecktor, and Oren Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 55–60, Ann Arbor, MI.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors. 2003. *Proceedings of the ACL 2003 workshop on multiword expressions: Analysis, acquisition and treatment*.

Chris Brockett. 2007. Aligning the RTE 2006 corpus. Technical Report MSR-TR-2007-77, Microsoft Research.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Aljoscha Burchardt, Nils Reiter, Stefan Thater, and Anette Frank. 2007. A semantic approach to textual entailment: System evaluation and task analysis. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 10–15, Prague, Czech Republic.

Michael Collins. 2002. Discriminative training methods for hidden Markov models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alch Buc, editors, *Machine Learning Challenges. Lecture Notes in Computer Science, Vol. 3944*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Trond Grenager, Bill MacCartney, Daniel Cer, Daniel Ramage, Chloé Kiddon, and Christopher D. Manning. 2007. Aligning semantic graphs for textual inference and machine reading. In *Proceedings of the AAAI Spring Symposium*.

Georgiana Dinu and Rui Wang. 2009. Inference rules and their application to recognizing textual entailment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 211–219, Athens, Greece.

Nicole Grégoire, Stefan Evert, and Su Nam Kim, editors. 2007. *Proceedings of the ACL workshop: A broader perspective on multiword expressions*.

Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176, Prague, Czech Republic.

Adrian Iftene and Alexandra Balahur-Dobrescu. 2007. Hypothesis transformation and semantic variability rules used in recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 125–130, Prague, Czech Republic.

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

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 768–774, Montréal, Canada.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of NAACL*.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii.

Erwin Marsi, Emiel Krahmer, and Wauter Bosma. 2007. Dependency-based paraphrasing for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 83–88, Prague, Czech Republic.

Begona Villada Moirón, Aline Villavicencio, Diana McCarthy, Stefan Evert, and Suzanne Stevenson, editors. 2006. *Proceedings of the ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multi-word expressions: a pain in the neck for NLP. In *Proceedings of CICLing*.

Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors. 2004. *Proceedings of the second ACL workshop on multiword expressions: Integrating processing*.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. 2007. Shallow semantic in fast textual entailment rule learners. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 72–77, Prague, Czech Republic.