Aligning Predicate-Argument Structures for Paraphrase Fragment Extraction

Michaela Regneri[*], **Rui Wang**[†], **Manfred Pinkal**[◇]
[*◇] Dept. of Computational Linguistics, Saarland University, Saarbrücken, Germany
[†]Language Technology Lab, DFKI Gmbh, Saarbrücken, Germany
regneri@coli.uni-saarland.de, ruiwang@dfki.de, pinkal@coli.uni-saarland.de

Abstract

Paraphrases and paraphrasing algorithms have been found of great importance in various natural language processing tasks. While most paraphrase extraction approaches extract equivalent sentences, sentences are an inconvenient unit for further processing, because they are too specific, and often not exact paraphrases. Paraphrase fragment extraction is a technique that post-processes sentential paraphrases and prunes them to more convenient phrase-level units. We present a new approach that uses semantic roles to extract paraphrase fragments from sentence pairs that share semantic content to varying degrees, including full paraphrases. In contrast to previous systems, the use of semantic parses allows for extracting paraphrases with high wording variance and different syntactic categories. The approach is tested on four different input corpora and compared to two previous systems for extracting paraphrase fragments. Our system finds three times as many good paraphrase fragments per sentence pair as the baselines, and at the same time outputs 30% fewer unrelated fragment pairs.

Keywords: paraphrasing, paraphrase fragments, semantic roles

1. Introduction

The recognition and extraction of paraphrases is a core task of natural language understanding, which is challenging but practically very useful. Applications of paraphrase resources and algorithms include document summarization (Barzilay et al., 1999), machine translation (Zhao et al., 2008a; Marton et al., 2009), natural language generation (Zhao et al., 2010; Ganitkevitch et al., 2011), plagiarism detection (Potthast et al., 2012), and recognizing textual entailment (Bosma and Callison-Burch, 2007).

One common way to create paraphrase resources is by extracting equivalent sentences from monolingual comparable corpora. However, sentences are often an impractical unit to use, as in the following example:

(1) *The patient gets out of bed and finds a pair of forceps to* ***extract his sore tooth*** *and rips it out of his mouth.*

(2) *Once the nurse leaves, he grabs a clamp and* ***pulls out the tooth that's hurting him****.*

While both sentences describe the same main event, they are clearly not exact paraphrases. The first sentence additionally describes an action that is omitted in the second sentence (*gets out of bed*), and the second sentence additionally mentions that *the nurse leaves*. However, there are still important parts of those sentences that should and can be matched as paraphrases, such as the two highlighted phrases.

Sentence pairs with similar meanings are easier to extract than smaller units that are exact paraphrases. In fact, it is hard to find two sentences in a parallel corpus which convey exactly the same information. Furthermore, sentence-sized units are hardly usable by other NLP applications or systems. A similar problem is known from machine translation: aligned sentence pairs do not help for the translation of unseen sentences, but just having aligned words (or a bilingual dictionary) provides too little context for capturing many linguistic phenomena and language variations.

With similar techniques used to create phrase tables for machine translation systems, there have been some recent approaches to extract *paraphrase fragments* from sentences with similar meaning (Bannard and Callison-Burch, 2005; Zhao et al., 2008b; Max, 2009; Wang and Callison-Burch, 2011; Regneri and Wang, 2012).

Previous work on fragment extraction already showed that shallow word-matching approaches are not sufficient for extracting such paraphrases, especially from input sentence pairs with only little word overlap. The linguistic knowledge put into such algorithms includes part-of-speech (POS) tags and chunk information (Barzilay and McKeown, 2001) as well as syntactic analysis via constituent trees (Callison-Burch, 2008; Zhao et al., 2008b) or dependency trees (Regneri and Wang, 2012).

Deeper semantic information has been mostly neglected by the approaches mentioned above, despite the fact that paraphrasing is the task of finding *semantically* equivalent linguistic expressions. In this paper, we therefore investigate the impact of semantic information on paraphrase fragment extraction, focusing on automatically induced semantic role annotations.

Our approach has two main advantages over previous ones:

Paraphrases with different *syntactic categories*: By matching semantic predicates, our system can produce paraphrases of different syntactic categories, like the clause *what he decided* and the noun phrase *his decision*.

Paraphrases with high *lexical variance*: We can extract paraphrases from sentence pairs with little word overlap, and thus also obtain fragment pairs with high wording variance.

We evaluate our approach on four corpora of different domains and sentence complexities, including traditional newswire articles and short video descriptions of daily life events.

As supplementary data, we provide our system output and our gold standard with fine-grained annotations comprising paraphrase and entailment information. This set could be useful for both paraphrase-related research and work on recognizing textual entailment (Dagan et al., 2005).

The remainder of the paper is organized as follows: Section 2 reviews related work on paraphrase acquisition and generation; Section 3 describes the input data we use for our system, which is then described in Section 4. Section 5 outlines our system evaluation, followed by analysis of the results as well as comparison to previous approaches. Section 6 concludes the paper and points out several directions for future research.

2. Related Work

Paraphrase acquisition methods can be distinguished by the input data they use, and by the actual algorithm that extracts the paraphrases.

As far as data sources are concerned, there has been a lot of work with parallel or comparable corpora. Barzilay and McKeown (2001) use different English translations of the same novels, which can be considered as monolingual parallel corpora. Equivalent translations as bilingual parallel corpora are also used by Bannard and Callison-Burch (2005) and Zhao et al. (2008b), who take one language as the pivot and match two possible translations in the other languages as paraphrases if they share a common pivot phrase.

There is also work on comparable corpora and paraphrase fragment extraction (Quirk et al., 2004; Wang and Callison-Burch, 2011) using multiple newspaper reports of the same events as source corpora. In addition, "targeted" paraphrasing systems, e.g. for the medical domain (Deléger and Zweigenbaum, 2009) or specific geographic topics (Belz and Kow, 2010) either perform the experiments in one specific domain (i.e., medical) or one specific topic (i.e., British geography).

Some recent approaches use crowd-sourcing techniques to collect parallel corpora with descriptions of everyday tasks that are usually not found in standard texts (Regneri et al., 2010; Chen and Dolan, 2011). Burrows et al. (2013) extract passage-level paraphrases via crowdsourcing. Bouamor et al. (2012) and Max et al. (2012) conduct several experiments to compare the impact of different data sources on paraphrase acquisition, for English and French.

Apart from the nature of their input data, paraphrasing approaches also differ in their acquisition methods and granularity of the final output. Many deliver sentence-level paraphrases (Barzilay and McKeown, 2001; Barzilay and Lee, 2003; Dolan et al., 2004; Quirk et al., 2004), while recent studies place more emphasis on phrase-level paraphrase extraction (Bannard and Callison-Burch, 2005; Wang and Callison-Burch, 2011; Martzoukos and Monz, 2012), in order to increase the applicability of the collected paraphrase pairs as resources for other NLP tasks.

Semantically interchangeable patterns can also be viewed as paraphrase resources (Shinyama et al., 2002; Shinyama and Sekine, 2003). Each pattern contains one or more anchor slots, which are usually restricted to certain types (e.g. named entities, *NEs*), and the paraphrases establish specific relations between the slots. Recent work by Fujita et al. (2012) and Shima and Mitamura (2012) did bootstrapping of both such patterns and their instances in large corpora. Zhao et al. (2008b) use more generic slots with the same POS, and do not target specific relations between them. More abstract patterns (or so-called *inference rules*) (Lin and Pantel, 2001; Szpektor et al., 2004) further loosen constraints on the slots, but the extracted paraphrases are mostly verbal phrases.

Our approach produces more general paraphrase fragment pairs, without any restrictions to particular syntactic categories. We follow the line of research started by Callison-Burch (2008) and Regneri and Wang (2012), who work on syntactic dependency trees, but carry it to the next level: Instead of matching syntactic structures, we find paraphrases based on semantic dependencies.

3. Input Data

We aimed to keep our system as generic as possible, and, at the same time, to analyze how different data sources affect our results. As input, we used sentential paraphrases from four different corpora. This section introduces those corpora and describes how they differ with respect to their domain or genre, their paraphrase extraction methods, their paraphrase assignment reliability, and the complexity of their sentences. (Table 4 contains some example sentences from each corpus.)

The Microsoft Paraphrase Corpus (Dolan and Brockett, 2005, MSR) is often used as a benchmark for paraphrase classification systems. MSR is based on automatically clustered newspaper texts, which were subsequently manually filtered for paraphrases. Given that it is hand-annotated, we estimate its accuracy at 100%. For our experiments, we take randomly selected paraphrases from the MSR test set.

The Microsoft Video Description Corpus (Chen and Dolan, 2011, MSVD) contains descriptions for a collection of short videos; the videos display scenes from various domains. Each video is described by multiple one-sentence descriptions in several languages. Sentences that describe the same video are often paraphrases, but according to Chen & Dolan, only 60% of them are in fact semantically equivalent. We consider only the English descriptions from MSVD.

The TACoS Corpus (Regneri et al., 2013) contains textual descriptions of mid-length videos. The descriptions consist of multiple sentences that are temporally aligned with the source videos, such that one sentence covers one scene event. A by-product of this temporal alignment are sentential paraphrases: There are multiple descriptions for the same video, and thus also multiple sentences that describe the same video snippet. In this respect, the corpus is similar to MSVD, but given the detailed temporal alignment, the paraphrase accuracy is probably higher (but was unfortunately not numerically evaluated). We thus cannot give an exact precision value, but estimate it higher than MSVD, but clearly not perfect.

Corpus	*Paraphrase Extraction Method*	# pairs	*Precision*	*Words / Sentence*	*Dice*
MSR	automated clustering + manual annotation	5,801	1.00	19.61	0.52
MSVD	same video stimulus	449,026	0.60	8.49	0.42
TACOS	same video + timestamp-based alignment	48,260	$0.60 <> 0.90$	9.06	0.36
HOUSE	same video + structural / semantic alignment	5,693	0.79	15.45	0.33

Table 1: Corpus comparison; *# pairs* is the number of paraphrases. The precision for TACOS is estimated.

The "House" Corpus (Regneri and Wang, 2012) was automatically created from parallel monolingual texts. Like for MSVD and TACoS, the parallel texts describe the same video, but in this case the video is a whole episode of the TV show *House M.D.* The paraphrases were automatically extracted from the episode recaps, considering the parallel sequential structures of the texts. According to Regneri and Wang, the paraphrase assignment has a precision of 79%.

Table 1 summarizes different benchmarks of the corpora. MSR is the only corpus with hand-tagged paraphrase information. The other three corpora have in common that their source texts somehow describe the same video, but the descriptions differ in length and annotation, and they use different paraphrase extraction methods: The video descriptions in MSVD are single sentences, so different descriptions of the same video are matched. The sentence pairs for TACoS result from alignments with the source video based on timestamps, and the House paraphrases are automatically extracted using semantic similarity and the linear structure of the source recaps.

We try to estimate processing complexity for each corpus by measuring paraphrase reliability, average sentence length, and surface similarity of sentence pairs. The assumption is that very reliable, very short and very similar paraphrases are easier to process than less accurate, long sentence pairs with little word overlap.

Precision is the estimated number of correct paraphrases in the corpus, which is highest for the manually tagged MSR, and presumably lowest for MSVD and the automatically computed House paraphrases. *Words / sentence* describes the average sentence length. The sentences in MSR and House contain more than twice as many words as the other two corpora. This is probably due to the fact that House and MSR cover standard texts (either episode recaps or newspaper articles), while TACoS and MSVD resulted from crowdsourcing of focused video descriptions.

Dice shows the average word overlap of paraphrases in each corpus, computed with the Dice coefficient. In MSR, each paraphrase pair shares half of their vocabulary on average, whereas the paraphrases in the House corpus show larger lexical differences between the matched sentences. As a consequence, we consider House as the most challenging input corpus, because it contains very long sentences with little word overlap and only moderate precision.

4. Fragment Extraction with Semantic Roles

This section first describes how we we preprocessed our corpora, and then outlines our paraphrase fragment extraction algorithm.

4.1. Preprocessing

We use different preprocessing pipelines: The core part of our system first uses TreeTagger (Schmid, 1994) to assign POS-tags. For dependency parsing, we use MSTParser (McDonald et al., 2005), a graph-based state-of-the-art dependency parser with CoNNL-style output. We apply a role labeler by Zhang et al. (2008) for semantic parsing, adding PropBank and NomBank annotations to the sentences. The result is a flat semantic parse with predicates and their arguments assigned. Figure 1 shows two example sentences with their semantic argument links, marked with ARG0, ARG1, and ARGM. The role labeler is trained on the CoNLL 2008 shared task dataset, for which it achieved state-of-the-art performance (Surdeanu et al., 2008).

In addition to this main parsing pipeline, we also extract dependency trees with the Stanford parser (Klein and Manning, 2003). While we needed the MSTParser for compatibility with the role labeler, we use the more fine-grained Stanford dependencies to find seed *anchors* for each sentence pair, as described in the following section.

4.2. Fragment Extraction

Our algorithm extracts pairs of phrases from a pair of semantic predicate-argument structures, supported by their underlying syntactic parse trees. We first match equivalent arguments (*anchors*) from both sentences, and then extract predicate-argument pairs that contain those anchors as their arguments. The final paraphrases are textual realizations of the predicate argument pairs.

The algorithm is a parallel graph search, extracting the largest weakly connected components that span the anchors and their dominating predicates. The anchors that serve as input are synonymous word pairs or references to the same entities from the two sentences. In the following, we will abbreviate the semantic predicates with PRED and their arguments with ARG. We call groups of predicates and arguments PREDSETs.

For a sentential paraphrase pair (s_1, s_2), we extract paraphrase fragments with the following steps:

4.2.1. Finding anchors

The graph search starts from matched word node pairs we call *anchors*. An anchor node in s_1 always has an (presumably synonymous) aligned anchor in s_2. We compare two different methods for identifying such anchors: The basic approach simply uses word matching, extracting all pairs of lemma-equal words as potential anchors. In order to increase recall, we generalize this very strict method using WordNet (Fellbaum, 1998) combined with dependency information from the Stanford parses. For this second anchoring method, we match dependency triples rather than

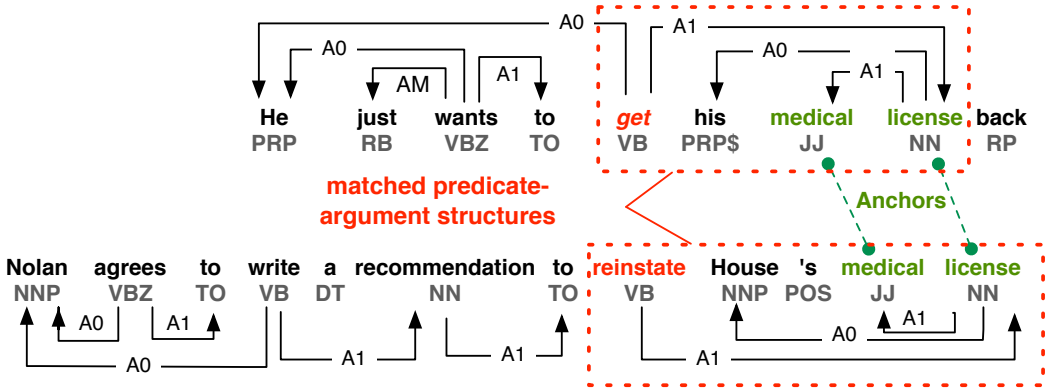

Figure 1: Example fragment extraction from two semantically parsed sentences.

words: Two dependency triples are matched if they have the same relation label and both of their arguments are synonyms according to WordNet. (We do not perform word sense disambiguation, but rather take words to be synonyms if they share at least one synset. Given the highly similar contexts in the sentential paraphrases, this crude heuristics does not noticeably affect the matching precision.) We consider the matched dependency arguments as anchor pairs. In Figure 1, we extracted two (adjacent) anchor pairs: the occurrences of *medical*, and the two instances of *license*.

4.2.2. Predicate-argument pair extraction

For each anchor found in the first step, we try to find a corresponding ARG. If the anchor is not an ARG itself, we recursively pick its syntactic head until we find one, or discard the anchor if none is found. In the example, all four anchors are arguments. We then group each anchored ARG with all its directly dominating predicates in a PREDSET. Note that there can be multiple predicates (dominating the same ARG) in one PREDSET. For the sentences in Figure 1, we would extract three PREDSETs for each sentence: $\{license_P, medical_{A1}\}$ for both sentences, plus $\{license_P, his_{A0}\}$ and $\{get_P, license_{A1}\}$ for the first one, and $\{reinstate_P, license_{A1}\}$ plus $\{license_P, House_{A0}\}$ for the second one.

4.2.3. Recursive fragment grouping

We join two PREDSETs P_i and P_j into one if there is at least one predicate in P_i that shares at least one argument with any predicate in P_j (the ARG does not need to be in the PREDSETs). We do this recursively until no intersecting PREDSETs are left. In our example, we collapse the PREDSETs with *license* as their predicate (which trivially shares all arguments). The system finally outputs two PREDSET pairs: $(\{license_P, medical_{A1} his_{A0}\}$, $\{license_P, medical_{A1}, House_{A0}\})$ and $(\{reinstate_P, license_{A1}\}$, $\{get_P, license_{A1}\})$. The output pairs are restricted to PREDSET pairs (P_1,P_2) if all anchors in P_1 are matched to anchors in P_2, and vice versa.

4.2.4. Surface realization

This step is necessary for manual evaluation, because the predicate-argument pairs are hard to interpret without context. We use the syntactic tree structures to extract the corresponding text span for each PREDSET: We take the smallest continuous text span that contains all elements of the PREDSET, and all their syntactic dependents. In the example, this would translate the PREDSET pair $(\{get, license\}, \{reinstate, license\})$ to the fragment pair *get his medical license back – reinstate House's medical license*.

5. Experiments

We evaluate the performance of our system on all input corpora and compare it to the baselines. Both system variants were tested separately: The first variant is the plain system using lemma matching to find anchors (*PaRole*), and the second variant additionally uses WordNet and dependency triples for anchor finding (*PaRole + WN*).

5.1. Gold Standard

For the final evaluation, we create a gold standard containing 100 paraphrase fragment pairs per input corpus and system configuration. We randomly sample 100 sentence pairs for each corpus and each system configuration, and then randomly pick one fragment pair per sentence pair. This results in a compilation of 800 fragment pairs.

Two annotators labelled each ordered fragment pair (F_1, F_2) with one of the following labels:

1. **paraphrases**: F_1 and F_2 are mutually exchangeable paraphrases.
2. **containment**: F_1 entails F_2, but also contains additional information.
3. **backwards containment**: F_2 entails F_1, but also contains additional information.
4. **related**: F_1 and F_2 have some core part in common, but neither of them fully entails the other.
5. **unrelated**: F_1 and F_2 share (almost) no content.
6. **invalid**: either F_1 or F_2 is completely ungrammatical or otherwise unreadable (e.g., *"n't"* as a single fragment).

We closely followed the annotation scheme proposed by Wang and Sporleder (2010) (originally for general textual semantic relations between sentences), leaving out the *contradiction* relation. The inclusion of the *entailment* labels

System	*Precision*		*relaxed Prec.*		*Productivity*		*pre * pro*		*rel * pro*	
	HOUSE	MSR	HOUSE	MSR	HOUSE	MSR	HOUSE	MSR	HOUSE	MSR
Giza-Baseline	0.28	**0.33**	0.57	0.64	0.76	0.33	**0.21**	0.11	0.43	0.21
VP-Baseline	**0.34**	n/a	0.84	n/a	0.42	n/a	0.14	n/a	0.35	n/a
PaRole	0.22	0.30	**0.94**	**0.94**	0.88	1.57	0.19	0.47	0.83	1.48
PaRole + WN	0.13	0.28	0.78	0.92	**1.54**	**1.82**	0.20	**0.51**	**1.20**	**1.67**

Table 2: Comparison of our two systems (last two rows) with two baselines on the HOUSE and the MSR corpus. (*pre * pro = precision * productivity*, *rel * pro = relaxed precision * productivity*)

allows us to explicitly distinguish fragments that are mutually equivalent (*paraphrases*) and those that are at least in some context exchangeable because one entails the other (*containment* / *backwards containment*).

The raters saw the original sentence pairs along with the fragment pairs for annotation. The overall rater agreement is $\kappa = 0.50$, according to Cohen's Kappa (*moderate agreement)*. Conflicts were resolved by a third annotator.

5.2. Metrics and Baselines

To evaluate our system, we compute precision, relaxed precision and productivity. Precision is computed in the standard fashion:

$$precision = \frac{\textit{no. of paraphrase pairs}}{\textit{no. of all fragment pairs}}$$

In previous work on paraphrase fragment extraction, most reported precision numbers are actually a measure we call *relaxed precision* (Callison-Burch, 2008; Wang and Callison-Burch, 2011; Regneri and Wang, 2012). According to this measure, any fragment pair that has a substantial overlap (i.e., it is not *unrelated* according to our annotation) is counted as a paraphrase. To compare our results to previous approaches, we thus compute *relaxed precision* as follows:

$$relaxed\ precision = \frac{\textit{no. of NOT unrelated pairs}}{\textit{no. of all fragment pairs}}$$

Intuitively, relaxed precision quantifies the amount of fragment pairs that are not complete nonsense, i.e., with some further processing, one could extract exact paraphrases out of them.

Recall is hard to measure for our approach, so we provide a quantitative evaluation by simply counting how many fragment pairs we can extract per input sentence pair. This number indicates how "productive" each approach is by fixing the input data size. We also provide two combined measures, *precision * productivity* and *relaxed precision * productivity*. They indicate how many good fragment pairs a system computes per sentence pair, either with respect to standard precision, or taking the relaxed version of it.

We compare our system's results with two previous approaches on paraphrase fragment extraction: The first one was introduced by Wang and Callison-Burch (2011, *Giza-Baseline*) and extracts fragments based on word-word alignments. Those alignments are computed with Giza++ (Och and Ney, 2003), and continuous spans of aligned words are extracted as fragment pairs. The second system was implemented by Regneri and Wang (2012, *VP-Baseline*) and simply matches verbal phrases of sentence pairs if they have some words in common, found either by Giza++ alignment or with plain string matching.

5.3. Results

Overall, 201 fragment pairs in our set are tagged as *paraphrase*, 148 with *containment*, 170 with *backwards containment*, 196 as *related*, 81 as *unrelated* and 5 as *invalid*. This results in a precision of 25%, whereas 89% of the fragment pairs we extracted have at least a significant overlap, and 65% stand in some entailment relation (either paraphrases or containment in either direction).

5.3.1. Comparison to the baselines

We compare our results to the baselines on the *House* and the *MSR* corpora. The performances of both baselines on the House corpus are reported by Regneri and Wang (2012); the results for the Giza-Baseline on MSR are taken from Wang and Callison-Burch (2011). We unfortunately do not have any results for the VP-Baseline on MSR.

House and MSR have different advantages for this comparison: MSR is a standard corpus for paraphrase classification tasks, which often servers as a benchmark for algorithms dealing with sentential paraphrases. House is the most complex corpus in our collection (cf. Table 3), probably due to its fairly long sentences, the little word overlap within paraphrases and the noise as by-product of automated paraphrase extraction. We thus provide a comparison on one standard corpus, and the most challenging corpus in our set.

Table 2 shows the results of our system configurations and the two baselines. Our approach is much more productive than both baselines, and still reasonably precise: Relaxed precision for our role matching approach is at 94% for both corpora, which beats all other systems. Considering only plain precision, we cannot beat the VP-Baseline on House or the Giza-Baseline on MSR, but on the other hand, our system produces far fewer unrelated paraphrases for both corpora.

Adding WordNet nearly doubles the productivity of the basic aligner on the House data, and produces over five times more fragments on MSR. The restrictive VP-Baseline delivers well below one third of those results.

Looking at the combinations of precision and productivity, we can see that our system configurations and the stronger Giza-Baseline are very close to each other on the House corpus, but we outperform the baseline by a large margin

	MSR		MSRVD		HOUSE		TACOS		AVERAGE	
	PaR	+WN	PaR	+WN	PaR	+WN	PaR	+WN	PaR	+WN
Precision	**0.30**	0.28	**0.24**	0.18	**0.22**	0.13	**0.35**	0.31	**0.28**	0.23
Relaxed Pre.	**0.94**	0.92	0.89	**0.90**	**0.94**	0.78	0.88	**0.90**	**0.91**	0.88
Productivity	1.57	**1.82**	0.60	**0.70**	0.92	**1.54**	0.57	**0.70**	0.92	**1.19**
$pre * pro$	0.47	**0.51**	**0.14**	0.13	**0.20**	**0.20**	0.20	**0.22**	0.25	**0.27**
$rel * pro$	1.48	**1.67**	0.53	**0.63**	0.86	**1.20**	0.50	**0.63**	0.83	**1.04**

Table 3: Results for the two system configurations, grouped by input corpus. (*PaR* = *PaRole*)

on MSR. Considering relaxed precision, our approach is clearly more productive than both baselines.

5.3.2. Comparison of different inputs

Table 3 compares our systems' results for different source corpora, with the last column showing the averages over all four. For each corpus, the best results are marked in boldface.

First, there are clear differences between our two system variants. While the strict anchor matching (*PaRole*, here *PR)* is usually more precise, using WordNet highly increases productivity (by 0.27 on average), with moderate loss of precision (probably due to the dependency-backup). As a consequence, the combined measures for WordNet anchoring are slightly better considering strict precision (+0.02), and considerably better with the relaxed precision combination (+0.21).

We also see big differences for the different source corpora. The House corpus leads to the worst precision, which confirms our intuition that this corpus is complicated to process. The short descriptions from TACoS lead to more accurate paraphrases: manual inspection reveals that many of the sentence pairs are already short and accurate paraphrases, so the extraction step often simply returns the original sentence pair. MSR and MSVD end up somewhere in between, whereas the performance on MSR is much better than on MSVD, probably due to MSR's high word overlap and its high precision. Despite the mixed quality of the input corpora, our system consistently manages to filter out many *unrelated* sentence pairs.

The big differences in productivity are mainly due to the role labeler's performance: The crowdsourced data from MSVD is partially ungrammatical, and the House summaries contain many rare words. As a result, we get fewer and less reliable outputs from the semantic parser.

5.3.3. Sample Output & Errors

Table 4 shows 8 examples from our gold standard, which are all good paraphrases. The lexical variance reflects the variance in the input sentence pairs, e.g. the fragments for House have an average dice score of 0.37, and the source sentence have 0.33 (cf. Table 1). We also can match support verb constructions like *doing somersaults* to *flipping over*, and fragments of different syntactic categories, like *what Tucker decided* and *Tucker's decision*.

The extracted paraphrases vary in their length; the smallest possible unit is a phrase including a predicate from the role labeler, like *her face*. In the extreme case, the algorithm can also return the initial sentence pair (we do not show such an example here). In many cases, we get several output fragments per sentence, and most logically, there are more and longer outputs for longer sentences.

Of course we also get some errors, partially due to the nature of the approach: we rely on the sentences to be semantically closely related. If two sentences actually describe different events but contain synonyms (which are matched as anchors), we will still compute fragment pairs like *a vascular problem* and *they never had such problems*. This problem is more prevalent when we use WordNet to relax the anchor matching, but equal person names (which happens often in the House corpus) and pronouns (we do not perform coreference resolution) cannot be handled by either approach. Given the results, we found it nevertheless reasonable not to constrain the paraphrase matching further, because we wanted to retain the systems' capability to match paraphrases with high lexical differences and different categories.

Like most paraphrase fragment extraction algorithms, our system often does not succeed at matching exactly the right text spans, resulting in "containment" or "related" cases. This is mostly due to our syntax-dependent surface realization, and to the precision of the semantic parser (which often simply misses out on predicates). This results e.g. in pairs like *gain an average of 11 years of life , free of cardiovascular disease* and *Still , the scientists said , a third of those taking it would benefit , gaining an average of 11 years free of cardiovascular disease.* In future work, this could be tackled by a better role labeler, and possibly by a direct generation of text from the predicates (without including the actual sentence material).

5.4. Comparison to bilingual approaches

Some *bilingual* paraphrasing systems are very similar to our approach but not directly comparable, because we used monolingual corpora exclusively.

Callison-Burch (2008) (including an earlier approach by Bannard and Callison-Burch (2005)) generated paraphrase fragment pairs from the Europarl corpus (Koehn, 2005). The productivity cannot be easily calculated, as they use sentence pairs from multiple language pairs. As for precision, we find their criteria similar to ours. Although they had finer-grained evaluation metrics of *meaning* and *grammaticality*, our relaxed precision can be viewed as a simplified version of their *meaning*. They achieve 0.61 for meaning only and 0.55 for both, when forcing the output fragment pair to share the same syntactic category, and 0.56 and 0.3 respectively without a such constraint. Both approaches generate a similar amount of paraphrases.

Source	Sentence 1	Sentence 2	Fragment 1	Fragment 2
HOUSE	The patient gets out of bed and finds a pair of forceps to extract his sore tooth and rips it out of his mouth.	Once the nurse leaves, he grabs a clamp and pulls out the tooth that's hurting him.	extract his sore tooth	pulls out the tooth that's hurting him
	Wilson tells House what Tucker decided, and House points out again that Tucker is a self-important jerk.	Later, Wilson tells House about Tucker's decision and admits he's a little disappointed.	what Tucker decided	Tucker 's decision
MSR	Prosecutors maintained that Durst murdered Black to try to assume Black's identity.	Prosecutors called Durst a cold-blooded killer who shot Black to steal his identity.	assume Black 's identity	steal his identity
	She said the president's eyes filled with tears when she told him he would have to confess to their teenage daughter as well.	Mrs Clinton writes her husband's eyes filled with tears when she told him he would have to confess to Chelsea as well.	have to confess to their teenage daughter as well	have to confess to Chelsea as well
TACOS	Girl throws away top of carrot	She discards the tops of the carrots.	top of carrot	the tops of the carrots
	She throws away the carrot peelings.	She discards the unwanted carrot shavings.	the carrot peelings	the unwanted carrot shavings
MSVD	A girl is doing make-up on her cheeks and forehead with a brush.	A woman is applying makeup to her face.	her cheeks and forehead	her face
	A cat is doing a somersault.	A cat is flipping over	doing a somersault	flipping over

Table 4: Examples of extracted paraphrase fragments, with their input corpora and source sentences.

Another related approach by Zhao et al. (2008b) extracted one paraphrase pattern from two sentence pairs on average, and achieved 0.67 precision. One major difference to our system is that they extracted *patterns* (instead of fragments) by replacing the anchor words with their POS tags, which also restricted the output fragment pairs to (partially) share the same syntactic categories.

6. Conclusion and Future Work

We presented a new approach for paraphrase fragment extraction using semantic roles from sentential paraphrases or near-paraphrases. We tested the algorithm on multiple corpora, and found that it is more productive than previous systems and outputs far fewer unrelated fragment pairs. We will distribute our gold standard with annotations of fine-grained categories as supplementary data.

For future work, we plan to refine the granularity of the extracted fragments, in order to find equivalents for phrases that are not predicates in the semantic parse. Motivated by the close relationship between paraphrase acquisition and recognizing textual entailment (Androutsopoulos and Malakasiotis, 2010), we also want to use our fine-grained gold standard to investigate the entailment relationship between paraphrase fragments and find a way to separate exact paraphrases from overlapping or entailing phrases.

7. Acknowledgments

The first author was funded by the Cluster of Excellence "Multimodal Computing and Interaction" in the German Excellence Initiative. The second Author was funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 287923 (EXCITEMENT, http://www.excitement-project.eu/). – We want to thank Jonas Sunde, Noushin Fadaei and David Przybilla for the extensive data annotation. We are especially grateful to Alexis Palmer and the anonymous reviewers for their helpful comments on previous versions of this paper. All remaining errors are of course our own.

8. References

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res.*, 38(1).

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proc. of ACL 2005*.

Barzilay, R. and Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proc. of HLT-NAACL 2003*.

Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proc. of ACL 2001*.

Barzilay, R., McKeown, K., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proc. of ACL 1999*.

Belz, A. and Kow, E. (2010). Extracting parallel fragments from comparable corpora for data-to-text generation. In *Proc. of INLG 2010*.

Bosma, W. and Callison-Burch, C. (2007). Paraphrase substitution for recognizing textual entailment. In *Proc. of CLEF 2006*.

Bouamor, H., Max, A., and Vilnat, A. (2012). Validation of sub-sentential paraphrases acquired from parallel monolingual corpora. In *Proc. of EACL 2012*.

Burrows, S., Potthast, M., and Stein, B. (2013). Paraphrase Acquisition via Crowdsourcing and Machine Learning. *Transactions on Intelligent Systems and Technology (ACM TIST)*.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proc. of EMNLP 2008*.

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proc. of ACL 2011*, Portland, OR, June.

Dagan, I., Glickman, O., and Magnini, B. (2005). The pascal recognising textual entailment challenge. In *MLCW*, pages 177–190.

Deléger, L. and Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proc. of the ACL-IJCNLP BUCC 2009 Workshop*.

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proc. of the third International Workshop on Paraphrasing*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proc. of COLING 2004*.

Fellbaum, C. (1998). *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.

Fujita, A., Isabelle, P., and Kuhn, R. (2012). Enlarging paraphrase collections through generalization and instantiation. In *Proc. of EMNLP-CoNLL 2012*.

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proc. of EMNLP 2011*.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proc. of ACL 2003*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of the tenth MT Summit*.

Lin, D. and Pantel, P. (2001). DIRT - Discovery of Inference Rules from Text. In *Proc. of the ACM SIGKDD 2001*.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In *Proc. of EMNLP 2009*.

Martzoukos, S. and Monz, C. (2012). Power-law distributions for paraphrases extracted from bilingual corpora. In *Proc. of EACL 2012*.

Max, A., Bouamor, H., and Vilnat, A. (2012). Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *Proc. of EMNLP-CoNLL 2012*.

Max, A. (2009). Sub-sentential paraphrasing by contextual pivot translation. In *Proc. of the 2009 Workshop on Applied Textual Inference*.

McDonald, R., Pereira, F., Ribarov, K., and Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proc. of HLT-EMNLP 2005*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In *Proc. of CLEF 2012*.

Quirk, C., Brockett, C., and Dolan, W. B. (2004). Monolingual machine translation for paraphrase generation. In *Proc. of EMNLP 2004*.

Regneri, M. and Wang, R. (2012). Using discourse information for paraphrase extraction. In *Proc. of EMNLP-CoNNL 2012*.

Regneri, M., Koller, A., and Pinkal, M. (2010). Learning Script Knowledge with Web Experiments. In *Proc. of ACL 2010*.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, (to appear).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc. of International Conference on New Methods in Language Processing*.

Shima, H. and Mitamura, T. (2012). Diversifiable bootstrapping for acquiring high-coverage paraphrase resource. In *Proc. of LREC 2012*.

Shinyama, Y. and Sekine, S. (2003). Paraphrase acquisition for information extraction. In *Proc. of the ACL Paraphrase 2003 Workshop*.

Shinyama, Y., Sekine, S., and Sudo, K. (2002). Automatic paraphrase acquisition from news articles. In *Proc. of HLT 2002*.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proc. of CoNLL 2008*.

Szpektor, I., Tanev, H., Dagan, I., and Coppola, B. (2004). Scaling Web-based Acquisition of Entailment Relations. In *Proc. of EMNLP 2004*.

Wang, R. and Callison-Burch, C. (2011). Paraphrase fragment extraction from monolingual comparable corpora. In *Proc. of the ACL BUCC 2011 Workshop*.

Wang, R. and Sporleder, C. (2010). Constructing a textual semantic relation corpus using a discourse treebank. In *Proc. of LREC 2010*.

Zhang, Y., Wang, R., and Uszkoreit, H. (2008). Hybrid learning of dependency structures from heterogeneous linguistic resources. In *Proc. of CoNLL 2008*.

Zhao, S., Niu, C., Zhou, M., Liu, T., and Li, S. (2008a). Combining multiple resources to improve smt-based paraphrasing model. In *Proc. of ACL-HLT 2008*.

Zhao, S., Wang, H., Liu, T., and Li, S. (2008b). Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. In *Proc. of ACL-HLT 2008*.

Zhao, S., Wang, H., Lan, X., and Liu, T. (2010). Leveraging Multiple MT Engines for Paraphrase Generation. In *Proc. of COLING 2010*.