

CONTENTS

<i>Oren Glickman & Ido Dagan</i>	
Acquiring lexical paraphrases from a single corpus	1
Index of Subjects and Terms	11

Acquiring Lexical Paraphrases from a Single Corpus

OREN GLICKMAN & IDO DAGAN

Bar Ilan University

Abstract

This paper studies the potential of extracting lexical paraphrases from a single corpus, focusing on the extraction of verb paraphrases. Most previous approaches detect individual paraphrase instances within a pair (or set) of comparable corpora, each of them containing roughly the same information, and rely on the substantial level of correspondence of such corpora. We present a novel method that successfully detects isolated paraphrase instances within a single corpus without relying on any a-priori structure and information.

1 Introduction

The importance of paraphrases has been recently receiving growing attention. Broadly speaking, paraphrases capture core aspects of variability in language, by representing (possibly partial) equivalencies between different expressions that correspond to the same meaning. Representing and tracking language variability is critical for many applications (Jacquemin 1999). For example, a question might use certain words and expressions while the answer, to be found in a corpus, might include paraphrases of the same expressions (Hermjakob et al. 2002). Another example is multi-document summarization (Barzilay et al. 1999). In this case, the system has to deduce that different expressions found in several documents express the same meaning; hence only one of them should be included in the final summary.

Recently, several works addressed the task of acquiring paraphrases (semi-) automatically from corpora. Most attempts were based on identifying corresponding sentences in parallel or ‘comparable’ corpora, where each corpus is known to include texts that largely correspond to texts in another corpus (see next section). The major types of comparable corpora are different translations of the same text, and multiple news sources that overlap largely in the stories that they cover. Typically, such methods first identify pairs (or sets) of larger contexts that correspond to each other, such as corresponding documents, by using clustering or similarity measures at the document level, and by utilizing external information such as requiring that corresponding documents will be from the same date. Then, within the corresponding contexts, the algorithm detects individual pairs (or sets) of sentences that largely overlap in their content and are thus assumed to describe the same fact or event.

Lin & Pantel (2001) propose a different approach for extracting ‘inference rules’, which largely correspond to paraphrase patterns. Their method extracts such paraphrases from a single corpus rather than from a comparable set of corpora. It is based on vector-based similarity (Lin 1998), which compares typical contexts in a global manner rather than identifying all actual paraphrase instances that describe the same fact or event.

The goal of our research is to explore further the potential of learning paraphrases within a single corpus. Clearly, requiring a pair (or set) of comparable corpora is a disadvantage, since such corpora do not exist for all domains, and are substantially harder to assemble. On the other hand, the approach of detecting actual paraphrase instances seems to have high potential for extracting reliable paraphrase patterns. We therefore developed a method that detects concrete paraphrase instances within a single corpus. Such paraphrase instances can be found since a coherent domain corpus is likely to include repeated references to the same concrete facts or events, even though they might be found within generally different stories. The first version of our algorithm was restricted to identify lexical paraphrases of verbs, in order to study whether the approach as a whole is at all feasible. The challenge addressed by our algorithm is to identify isolated paraphrase instances that describe the *same* fact within a single corpus. Such paraphrase instances need to be distinguished from instances of *distinct* facts that are described in similar terms. These goals are achieved through a combination of statistical and linguistic filters and a probabilistically motivated paraphrase likelihood measure. We found that the algorithmic computation needed for detecting such local paraphrase instances across a single corpus should be quite different than previous methods developed for comparable corpora, which largely relied on a-priori knowledge about the correspondence between the different stories from which the paraphrase instances are extracted.

We have further compared our method to the vector-based approach of (Lin & Pantel 2001). The precision of the two methods on common verbs was comparable, but they exhibit some different behaviors. In particular, our instance-based approach seems to help assessing the reliability of candidate paraphrases, which is more difficult to assess by global similarity measures such as the measure of Lin and Pantel.

2 Background and related work

The importance of modeling semantic variability has been recently receiving growing attention. Dagan & Glickman (2004) propose a generic framework for modeling textual entailment that recognizes language variability at a shallow semantic level and relies on a knowledge base of paraphrase patterns. Consequently, acquisition of such paraphrase patterns is of great significance. Lexical resources such as WordNet are commonly utilized, however they tend to be too

broad and do not contain the necessary domain vocabulary. Similarity-based lexical approaches (Lin 1998) are also inappropriate for semantic entailment for they do not capture equivalence and entailment of meaning but rather broader meaning similarity. Several works addressed the task of automatically acquiring paraphrase patterns from corpora. (Barzilay & McKeown 2001) use sentence alignment to identify paraphrases from a corpus of multiple English translations of the same text. (Pang et al. 2003) also use a parallel corpus of Chinese-English translations to build finite state automata for paraphrase patterns, based on syntactic alignment of corresponding sentences. (Shinyama et al. 2002) learn structural paraphrase templates for Information extraction from a comparable corpus of news articles from different news sources over a common period of time. Similar news article pairs from the different news sources are identified based on document similarity. Sentence pairs are then identified based on the overlap of Named Entities in the matching sentences. (Barzilay and Lee 2003) also utilizes a comparable corpus of news articles to learn paraphrase patterns, which are represented by word lattice pairs. Patterns originating from the same day but from different newswire agencies are matched based on entity overlap. We compare our results to those of the algorithm by (Lin & Pantel 2001), which extracts paraphrase-like inference rules for question answering from a single source corpus. The underlying assumption in their work is that paths in dependency trees that connect similar syntactic arguments (slots) are close in meaning. Rather than considering a single feature vector that originates from the arguments in both slots, vector-based similarity was computed separately for each slot. The similarity of a pair of binary paths was defined as the geometric mean of the similarity values that were computed for each of the two slots.

3 Algorithm

Our proposed algorithm identifies candidates of corresponding verb paraphrases within pairs of sentences. We define a *verb instance pair* as a pair of occurrences of two distinct verbs in the corpus. A *verb type pair* is a pair of verbs detected as a candidate lexical paraphrase.

3.1 *Preprocessing and representation*

Our algorithm relies on a syntactic parser to identify the syntactic structure of the corpus sentences, and to identify verb instances. We treat the corpus uniformly as a set of distinct sentences, regardless of the document or paragraph they belong to. For each verb instance we extract the various syntactic components that are related directly to the verb in the parse tree. For each such component we extract its lemmatized head, which is possibly extended to capture a semantically specified constituent. We extended the heads with any lexical modifiers that constitute a multi-word term, noun-noun modifiers, numbers and prepositional ‘of’

subject	secretary_general_boutros_boutros_ghali	subject	iraqi_force
object	implementation_of_deal	object	kurdish_rebel
modifier	after	pp-on	august_31
(A) verb:	delay	(B) verb:	attack

Fig. 1: *Extracted verb instances for sentence “But U.N. Secretary-General Boutros Boutros-Ghali delayed implementation of the deal after Iraqi forces attacked Kurdish rebels on August 31.”*

complements. Verb instances are represented by the vector of syntactic modifiers and their lemmatized fillers. For illustration, Figure 1 shows an example sentence and the vector representations for its two verb instances.

3.2 Identifying candidate verb instance pairs (filtering)

We apply various filters in order to verify that two verb instances are likely to be paraphrases describing the same event. This is an essential part of the algorithm since we do not rely on the high a-priori likelihood for finding paraphrases in matching parts of comparable corpora.

We first limit our scope to pairs of verb instances that share a common (extended) subject and object which are not pronouns. Otherwise, if either the subject or object differ between the two verbs then they are not likely to refer to the same event in a manner that allows substituting one verb with the other. Additionally, we are interested in identifying sentence pairs with a significant overall term overlap, which further increases paraphrase likelihood for the same event. This is achieved with a standard (Information Retrieval style) vector-based approach, with tf-idf term weighting

- $tf(w) = freq(w)$ in sentence
- $idf(w) = \log(N/freq(w))$ in corpus) where N is the total number of tokens in the corpus.

Sentence overlap is measured simply as the dot product of the two vectors. We intentionally disregard any normalization factor (such as in the cosine measure) in order to assess the absolute degree of overlap, while allowing longer sentences to include also non-matching parts that might correspond to complementary aspects of the same event. Verb instance pairs whose sentence overlap is below a specified threshold are filtered out.

An additional assumption is that events have a unique propositional representation and hence verb instances with contradicting vectors are not likely to describe the same event. We therefore filter verb instance pairs with contradicting propositional information - a common syntactic relation with different arguments. As an example, the sentence “Iraqi forces captured Kurdish rebels on August 29.” Has a contradicting ‘on’ preposition argument with the sentence from Figure 1(B) (“August 29” vs. “August 31”).

3.3 Computing paraphrase score of verb instance pairs

Given a verb instance pair (after filtering), we want to estimate the likelihood that the two verb instances are paraphrases of the same fact or event. We thus assign a paraphrase likelihood score for a given verb instance pair I_{v_1, v_2} , which corresponds to instances of the verb types v_1 and v_2 with overlapping syntactic components p_1, p_2, \dots, p_n . The score corresponds (inversely) to the estimated probability that such overlap had occurred by chance in the entire corpus, capturing the view that a low overlap probability (i.e., low probability that the overlap is due to chance) correlates with paraphrase likelihood. We estimate the overlap probability by assuming independence of the verb and each of its syntactic components as follows:

$$\begin{aligned} P(I_{v_1, v_2}) &= P(\text{overlap}) = P(v_1, p_1, \dots, p_n) P(v_2, p_1, \dots, p_n) \\ &= P(v_1) P(v_2) \prod_{i=1}^n P(p_i)^2 \end{aligned} \quad (1)$$

Where the probabilities were calculated using Maximum Likelihood estimates based on the verb and argument frequencies in the corpus.

3.4 Computing paraphrase score for verb type pairs

When computing the score for a verb type pair we would like to accumulate the evidence from its corresponding verb instance pairs. Following the vein of the previous section we try to estimate the joint probability that these different instance pairs occurred by chance. Assuming instance independence, we would like to multiply the overlap probabilities obtained for all instances. We have found, though, that verb instance pairs whose two verbs share the same subject and object are far from being independent (there is a higher likelihood to obtain additional instances with the same subject-object combination). To avoid complex modeling of such dependencies we picked only one verb instance pair for each subject-object combination, taking the one with lowest probability (highest score). This yields the set $T(v_1, v_2) = (I_1, \dots, I_n)$ of best scoring (lowest probability) instances for each distinct subject and object components. Assuming independence of occurrence probability of these instances, we estimate the probability $P(T(v_1, v_2)) = \prod P(I_i)$, where $P(I)$ is calculated by Equation (1) above. The score of a verb type pair is given by: $\text{score}(v_1, v_2) = -\log P(T(v_1, v_2))$.

4 Evaluation and analysis

4.1 Setting

We ran our experiments on the first 15-million word (token) subset of the Reuters Corpus. The corpus sentences were parsed using the Minipar dependency parser

(Lin 1993). 6,120 verb instance pairs passed filtering (with overlap threshold set to 100). These verb instance pairs derive 646 distinct verb type pairs, which were proposed as candidate lexical paraphrases along with their corresponding paraphrase score.

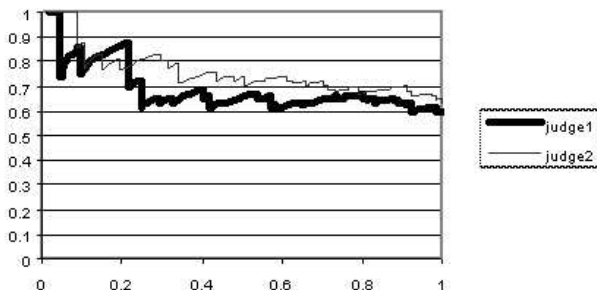


Fig. 2: Precision (y axis) recall (x axis) curves of system paraphrases by judge (verb type pairs sorted by system score)

The correctness of the extracted verb type pairs was evaluated over a sample of 215 pairs (one third of the complete set) by two human judges, where each judge evaluated one half of the sample. In a similar vein to related work in this area, judges were instructed to evaluate a verb type pair as a *correct* paraphrase only if the following condition holds: one of the two verbs can replace the other within some sentences such that the meaning of the resulting sentence will entail the meaning of the original one. To assist the judges in assessing a given verb type pair they were presented with example sentences from the corpus that include some matching contexts for the two verbs (e.g., sentences in which both verbs have the same subject or object). Notice that the judgment criterion allows for directional paraphrases, such as $\langle \text{invade}, \text{enter} \rangle$ or $\langle \text{slaughter}, \text{kill} \rangle$, where the meaning of one verb entails the meaning of the other, but not vice versa.

4.2 Results of the paraphrase identification algorithm

Figure 2 shows the precision vs. recall results for each judge over the given test-sets. The evaluation was conducted separately also by the authors on the full set of 646 verb pairs, obtaining comparable results to the independent evaluators. In terms of agreement, the Kappa value (measuring pair wise agreement discounting chance occurrences) between the authors and the independent evaluators' judgments were 0.61 and 0.63, which correspond to a substantial agreement level (Landis & Koch 1977). The overall precision for the complete test sample is 61.4% accuracy, with a confidence interval of [56.1,66.7] at the 0.05 significance level. Figure 3 shows the top 10 lexical paraphrases, and a sample

of the remaining ones, achieved by our system along with the annotators' judgments. Figure 4 shows correct sentence pairs describing a common event, which were identified by our system as candidate paraphrase instances.

1-	⟨fall, rise⟩	8+	⟨cut, lower⟩	302+	⟨kill, slaughter⟩
2+	⟨close, end⟩	9-	⟨rise, shed⟩	362+	⟨bring, take⟩
3+	⟨post, report⟩	10+	⟨fall, slip⟩	422+	⟨note, say⟩
4+	⟨recognize, recognize⟩	62+	⟨honor, honour⟩	482-	⟨export, load⟩
5+	⟨fire, launch⟩	122+	⟨advance, rise⟩	542+	⟨downgrade, relax⟩
6+	⟨drop, fall⟩	182+	⟨benefit, bolster⟩	602+	⟨create, establish⟩
7+	⟨regard, view⟩	242+	⟨approve, authorize⟩	632-	⟨announce, give⟩

Fig. 3: *Example of system output with judgments*

An analysis of the incorrect paraphrases showed that roughly one third of the errors captured verbs with contradicting semantics or antonyms (e.g., ⟨rise, fall⟩, ⟨buy, sell⟩, ⟨capture, evacuate⟩) and another third were verbs that tend to represent correlated events with strong semantic similarity (e.g., ⟨warn, attack⟩, ⟨reject, criticize⟩). These cases are indeed quite difficult to distinguish from true paraphrases since they tend to occur in a corpus with similar overlapping syntactic components and within quite similar sentences. Figure 4 also shows examples of misleading sentence pairs demonstrating the difficulties posed by such instances. It should be noticed that our evaluation was performed at the verb type level. We have not evaluated directly the correctness of the individual paraphrase instance pairs extracted by our method (i.e., whether the two instances in a paraphrase pair indeed refer to the same fact). Such evaluation is planned for future work. Finally, a general problematic (and rarely addressed) issue in this area of research is how to evaluate the coverage or recall of the extraction method relative to a given corpus.

4.3 *Comparison with (Lin & Pantel 2001)*

We applied the algorithm of (Lin & Pantel 2001), denoted here as the LP algorithm, and computed their similarity score for each pair of verb types in the corpus. To implement the method for lexical verb paraphrases, each verb type was considered as a distinct path whose subject and object play the roles of the X and Y slots.

As it turned out, the similarity score of LP does not behave uniformly across all verbs. For example, many of the top 20 highest scoring verb pairs are quite erroneous (see Figure 5), and do not constitute lexical paraphrases (compare with the top scoring verb pairs for our system in Figure 3). The similarity scores do seem meaningful within the context of a single verb v , such that when sorting all other verbs by the LP score of their similarity to v correct paraphrases are more likely to occur in the upper part of the list. Yet, we are not aware of a criterion

correct paraphrase instance pairs

Campbell is buying Erasco from Grand Metropolitan Plc of Britain for about \$210 million.	Campbell is purchasing Erasco from Grand Metropolitan for approximately US\$210 million.
The stock of Kellogg Co. dropped Thursday after the giant cereal maker warned that its earnings for the third quarter will be 20 percent below a year ago.	The stock of Kellogg Co. fell Thursday after it warned about lower earnings this year ...
Ieng Sary on Wednesday formally announced his split with top Khmer Rouge leader Pol Pot, and said he had formed a rival group called the Democratic National United Movement.	In his Wednesday announcement Ieng Sary, who was sentenced to death in absentia for his role in the Khmer Rouge's bloody rule, confirmed his split with paramount leader Pol Pot.

misleading instance pairs

Last Friday, the United States announced punitive charges against China's 1996 textile and apparel quotas ...	China on Saturday urged the United States to rescind punitive charges against Beijing's 1996 textile and apparel quotas ...
Municipal bond yields dropped as much as 15 basis points in the week ended Thursday, erasing increases from the week before.	Municipal bond yields jumped as much as 15 basis points over the week ended Thursday ...
Rand Financials notably bought October late while Chicago Corp and locals lifted December into by stops.	Rand Financials notably sold October late while locals pressured December.

Fig. 4: *Examples of instance pairs*

that predicts whether a certain verb has few good paraphrases, many or none. Given this behavior of the LP score we created a test sample for the LP algorithm by randomly selecting verb pairs of equivalent similarity rankings relative to the original test sample. Notice that this procedure is favorable to the LP method for it is evaluated at points (verb and rank) that were predicted by our method to correspond to a likely paraphrase.

The resulting 215 verb pairs were evaluated by the judges along with the sample for our method, while the judges did not know which system generated each pair. The overall precision on the LP method for the sample was 51.6%, with a confidence interval of [46.1,57.1] at the 0.05 significance level. The LP

<misread, misjudge>(0.62); <barricade, sandbag>(0.29); <disgust, mystify>(0.27); <jack, decontrol>(0.27); <Pollinate, pod>(0.25); <mark_down, decontrol>(0.23); <subsidize, subsidise>(0.22); <wake_up, divine>(0.21); <thrill, personify>(0.21); <mark_up, decontrol>(0.20); <flatten, steepen>(0.20); <mainline, pip>(0.20); <misinterpret, relive>(0.20); <remarry, flaunt>(0.19); <distance, dissociate>(0.18); <trumpet, drive_home>(0.18); <marshal, beleaguer>(0.17); <dwell_on, feed_on>(0.17); <scrutinize, misinterpret>(0.16); <disable, counsel>(0.16)
--

Fig. 5: *Top 20 verb pairs from similarity system*

results for this sample were thus about 10 points lower than the results for our comparable sample, but the two confidence intervals overlap slightly. It is interesting to note that the precision of the LP algorithm over all pairs of rank 1 was also 51%, demonstrating that just rank on its own is not a good basis for paraphrase likelihood. Figure 6 shows overall recall vs. precision from both

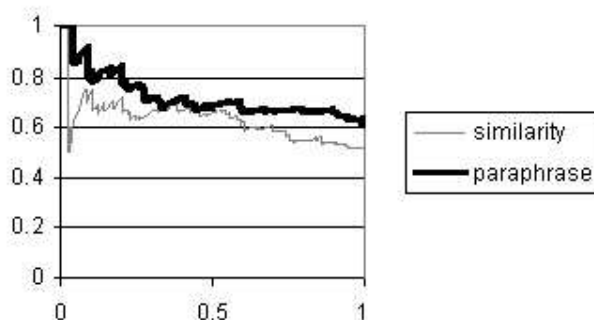


Fig. 6: Precision recall curve for our paraphrase method and LP similarity

judges for the two systems. The results above show that the precision of the vector-based LP method may be regarded as comparable to our instance-based method, in cases where one of the two verbs was identified by our method to have a corresponding number of paraphrases. The obtained level of accuracy for these cases is substantially higher than for the top scoring pairs by LP. This suggests that our approach can be combined with the vector-based approach to obtain higher reliability for verb pairs that were extracted from actual paraphrase instances.

5 Conclusions

This work presents an algorithm for extracting lexical verb paraphrases from a single corpus. To the best of our knowledge, this is the first attempt to identify actual paraphrase instances in a single corpus and to extract paraphrase patterns directly from them. The evaluation suggests that such an approach is indeed viable, based on algorithms that are geared to overcome many of the misleading cases that are typical for a single corpus (in comparison to comparable corpora). Furthermore, a preliminary comparison suggests that verb pairs extracted by our instance-based approach are more reliable than those based on global vector similarity. As a result, an instance-based approach may be combined with a vector-based approach in order to assess better the paraphrase likelihood for many verb pairs. Future research is planned to extend the approach to handle more complex

paraphrase structures and to increase its performance by relying on additional sources of evidence.

REFERENCES

- Barzilay, Regina & Lillian Lee. 2003. "Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment". *Proceedings of the Human Language Technology Conference (HLT-NAACL'03)*, 16-23. Edmonton, Canada.
- Barzilay, Regina & Kathleen McKeown. 2001. "Extracting Paraphrases from a Parallel Corpus". *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, 50-57. Toulouse, France.
- Barzilay, Regina, Kathleen McKeown & Michael Elhadad. 1999. "Information Fusion in the Context of Multidocument Summarization". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 550-557. Univ. of Maryland, College Park, Maryland, U.S.A.
- Dagan, Ido & Glickman Oren. 2004. "Probabilistic Textual Entailment: Generic Applied Modeling Of Language Variability". *PASCAL workshop on Text Understanding and Mining*, Grenoble, France.
- Hermjakob, Ulf, Abdessamad Echihabi & Daniel Marcu. 2002. "Natural Language Based Reformulation Resource and Web Exploitation for Question Answering". *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 801-810. Gaithersburg, Maryland, U.S.A.
- Jacquemin, Christian. 1999. "Syntagmatic and Paradigmatic Representations of Term Variation". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 341-348. Univ. of Maryland, College Park, Maryland, U.S.A.
- Landis, J.R. & G.G. Koch. 1997. "The Measurements of Observer Agreement for Categorical Data". *Biometrics* 33:159-174.
- Lin, Dekang. 1993. "Principle-Based Parsing without Overgeneration". *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, 112-120. Columbus, Ohio.
- Lin, Dekang. 1998. "Automatic Retrieval and Clustering of Similar Words". *Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL'98)*, 768-774. Montreal, Canada.
- Lin, Dekang & Patrick Pantel. 2001. "Discovery of Inference Rules for Question Answering". *Natural Language Engineering* 7:4.343-360.
- Pang, Bo, Kevin Knight & Daniel Marcu. 2003. "Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences". *Proceedings of the Human Language Technology Conference (HLT-NAACL'03)*, 181-188. Edmonton, Canada.
- Shinyama, Yusuke, Satoshi Sekine, Kiyoshi Sudo & Ralph Grishman. 2002. "Automatic Paraphrase Acquisition from News Articles". *Proceedings of the Human Language Technology Conference (HLT'02)*, 51-58. San Diego, Calif., U.S.A.

Index of Subjects and Terms

C.

cosine measure 4

I.

information retrieval 4

M.

multi-document summarization 1

P.

paraphrase

 acquisition 1

 importance 1

 lexical 2

parser

 dependency 6

Q.

question answering (QA) 1

S.

similarity measures 1

T.

tf-idf 4