Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 89-95 recibido 27-01-12 revisado 09-03-12 aceptado 18-03-12

Tree Edit Distance as a Baseline Approach for Paraphrase Representation*

Distancia de edición de árboles como caso base para la representación de la paráfrasis

Marta Vila
Universitat de Barcelona
Gran Via 585
08007 Barcelona
marta.vila@ub.edu

Mark Dras
Macquarie University
Herring Rd, North Ryde
NSW 2109
mark.dras@mq.edu.au

Resumen: Encontrar un formalismo adecuado para representar la paráfrasis constituye un reto para el Procesamiento del Lenguaje Natural. En este artículo, se analiza la distancia de edición de árboles como caso base para dicha representación. Los experimentos realizados utilizando Edit Distance Textual Entailment Suite muestran que, dado que la distancia de edición de árboles es una aproximación puramente sintáctica, las paráfrasis no basadas en reorganizaciones estructurales no encuentran una representación adecuada. Asimismo, muestran la necesidad de mejorar la forma como los árboles se alinean.
Palabras clave: Paráfrasis, distancia de edición de árboles, alineación de árboles.

Abstract: Finding an adequate paraphrase representation formalism is a challenging issue in Natural Language Processing. In this paper, we analyse the performance of Tree Edit Distance as a paraphrase representation baseline. Our experiments using Edit Distance Textual Entailment Suite show that, as Tree Edit Distance consists of a purely syntactic approach, paraphrase alternations not based on structural reorganizations do not find an adequate representation. They also show that there is much scope for better modelling of the way trees are aligned.
Keywords: Paraphrasing, tree edit distance, tree alignment.

1 Introduction

In paraphrasing, different wordings express same meaning. For example, an active/passive voice alternation occurs in the paraphrase pair in (1).[1]

(1) a. The guide drew our attention to a [...] dungeon
b. Our attention was drawn by [the] guide to a [...] dungeon

String pairs like the one in (1) are obviously not very general. Formally representing paraphrasing, i.e., transforming paraphrase strings into paraphrase patterns by enriching them with linguistic knowledge and, at the same time, making them more general, makes paraphrase knowledge more efficient and scalable to various Natural Language Processing (NLP) tasks and applications. In (2), a representation of the active/passive alternation in (1) along the lines of the original Transformational Grammar representation of Chomsky (1957) can be observed. All linguistic units but prepositions have been substituted by the corresponding morpho-syntactic categories, which are mapped from one member of the pair to the other.

(2) a. NP_1 V_{active} NP_2 *to* NP_3.
b. NP_2 $V_{passive}$ *by* NP_1 *to* NP_3.

* We are grateful to M. Antònia Martí and Horacio Rodríguez for their helpful advice as experienced researchers. We would also like to express our gratitude to the anonymous reviewers for their suggestions to improve this article. This research work is supported by the TEXT-Knowledge 2.0 (TIN2009-13391-C04-04) MICINN project. Also, the work of the first author is financed by the FPU AP2008-02185 MECD grant and, within it, the funding for a 6-month stay in Macquarie University.

[1] Example from the P4P corpus. `http://clic.ub.edu/corpus/en/p4p`.

ISSN 1135-5948

Paraphrasing is a complex phenomenon, where many linguistic mechanisms—shallow or deep, formal or conceptual—can be displayed. Contrary to (1), in the example in (3),[2] a formal structural mapping between the two members of the pair in italics cannot be established.

(3) a. Michael Mitchell [...] *did not answer his phone* Wednesday afternoon
b. Michael Mitchell [...] *was not available for comment*

In this paper, we want to capture two things with respect to paraphrase representation. Primarily, we are interested in how well a representation can capture the mapping of structures (typically as instantiated by tree alignment) that occur in paraphrasing. By way of illustration, if the structural representation of (4-a) maps to the structural representation of (4-b),[3] and, in the former, *estimated* is the head of the dependent noun *people*, while the reverse is true in the latter (i.e., *people* is the head of *estimated*), the paraphrase representation must be able to capture that. That is, the trees should be aligned in a way that maps corresponding nodes to each other.

(4) a. It is estimated that 200,000 people are left behind
b. An estimated 200,000 people left behind

Secondarily, we want a representation approach capable of dealing with paraphrase complexity at a reasonable computational cost. The intrinsic variety of paraphrasing demands a highly expressive representation. Nevertheless, high expressive capacity generally entails low computational efficiency, as, in general, there is a trade-off between the two. Thus, finding an adequate balance is needed.

Given all of this, our first objective is to build a paraphrase representation baseline (in terms of expressiveness) to evaluate its level of coverage of the paraphrase phenomenon and the potential drawbacks that it presents in alignment. This baseline will be the basis for a further analysis of more expressive approaches.

Wu (2010) presents a general framework for considering alignment, including tree alignment, which is useful for considering a range of possible representations. In this paper, we work with Tree Edit Distance (TED). Given two dependency trees, TED allows for establishing the distance between them according to the number of edit operations on tree fragments (insertion, deletion, substitution) required to go from one to the other. It can thus convert any tree into any other arbitrary tree (e.g., by deleting all nodes and then inserting new nodes). We take these two dependency trees and the edit operations mapping between them as a paraphrase representation baseline, and investigate its performance regarding paraphrasing and how well the mapping of items that should be aligned is preserved.

We use the Edit Distance Textual Entailment Suite (EDITS),[4] a software package aimed at Recognizing Textual Entailment (RTE) relations between two portions of text, which embeds an implementation of the TED algorithm described in Shasha and Zhang (1990). Given that paraphrasing can be considered a bidirectional entailment, we hypothesize that such a package can also be used in the paraphrase domain. As will be seen, we do not use EDITS as a textual entailment or paraphrase classifier, and we focus instead on the edit operations between trees it provides.

Using EDITS, we represented a set of paraphrase pairs with the TED approach. First, we analyzed the coverage of this approach within the paraphrase phenomenon and the drawbacks that presents. We then proceeded to our main question of interest, how well the structures are mapped.

In what follows, after presenting a brief state of the art on paraphrase representation (Section 2), we set out our experiments and results (Section 3). Conclusions and future lines of research appear in Sections 4 and 5, respectively.

2 State of the Art

Choosing a paraphrase representation formalism implies seeking balance between expressivity and computational cost. The least

[2] Example from the MSRP corpus. `http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/`.

[3] Example from Wan (2010).

[4] `http://edits.fbk.eu/`

expressive way consists in simply stating the paraphrase nature of a pair of strings. An example of this approach is the Microsoft Research Paraphrase Corpus (MSRP) (Dolan and Brockett, 2005),[5] which consists of a set of sentence pairs with a yes/no paraphrase judgement.

Expressivity may be increased transforming pairs of strings into pairs of regular expression patterns, which can be synchronized using Finite State Transducers (FST) or their probabilistic version, namely, Stochastic Finite State Transducers (SFST). Casacuberta and Vidal (2007), for instance, learn finite-state models for Machine Translation (MT). Drawbacks facing FSTs are that they are too constrained to model paraphrase mapping complexity because they only compare regular expression strings (not deeper representations), and that their expressive power is limited to Regular Grammars (RG).

A further step in expressivity, but with a higher computational cost, may be the use of the bilingual version of Context Free Grammars, namely, Synchronous Context Free Grammars (SCFG) (Aho and Ullman, 1969), which simultaneously produce strings in two languages. Dekai Wu, starting in Wu (1997), proposed several subclasses of SCFGs pruning their expressivity for reducing their computational cost, e.g., Inversion Transduction Grammars (ITG), Bracketting ITG (BITG), Linear ITG (LITG), together with their probabilistic versions. Some of these formalisms are proved to be expressive enough for learning and parsing.

A richer level of expresivity may be found in the family of the tree transducers, within the framework of Mildly Context Sensitive Grammars (MCSG). Some examples are Quasi-Synchronous Grammars (QSG), which were proposed by Smith and Eisner (2006); Synchronous Tree Adjoining Grammars (STAG), which Dras (1999) applies to syntactic paraphrasing; or Synchronous Tree-Substitution Grammar (STSG) (Eisner, 2003), a restricted version of STAGs. All these proposals have been mainly applied to MT (translation between different languages), but they may also be applied to paraphrasing (understood as translation within the same language).

TED is the approach chosen in this paper as a baseline for paraphrase representation. There is work in the related field of RTE (Kouylekov and Magnini, 2005) and also in paraphrasing (Heilman and Smith, 2010), but there all that is of interest is optimising the mapping between strings, not between structures. That is, how the trees are transformed is unimportant in those applications, as long as the transformation of the string is carried out with a minimum cost. In contrast, our interest is precisely on the tree mapping.

3 Experiments and Results

We performed two different experiments aiming at the analysis of TED performance for paraphrase representation (Section 3.1) and the analysis of the problem of tree alignments (Section 3.2). In both of them, we used EDITS.

EDITS presents a modular structure, whose main components are: algorithms used to compute a distance score; cost schemes defining the cost for each edit operation; a cost optimizer, which adapts cost schemes to specific datasets; and rules providing linguistic knowledge. Using as a starting point these modules, plus a training corpus with sentence pairs annotated with yes/no textual entailment or paraphase judgement, EDITS builds a model, which will be subsequently used to classify unseen sentence pairs.

The EDITS output which we are interested in (considering we have selected, among the possible algorithms, the TED one) consists of a file with the dependency trees of each member of the pair,[6] and a file with the edit operations between them, a score and an entailment/paraphrase judgement. Our focus is on how the trees are transformed, i.e., the trees and edit operations, not on the final score nor classification.

3.1 The Performance of TED for Paraphrase Representation

Our objective here is analyzing a set of paraphrase pairs with EDITS to see the coverage of TED regarding the paraphrase phenomenon and the potential problems arisen. The corpus used is the MSRP,[7] because it is a reference corpus fulfilling EDITS requirements: it contains a large quantity of data

[5] See note 2.

[6] The Stanford parser is the one used by EDITS. `http://nlp.stanford.edu/software/lex-parser.shtml`

[7] See Section 2 for references.

(5,800 English sentence pairs) with manual annotations indicating whether they are paraphrases (67%) or not (33%). It is already divided into training (70%) and test (30%).

We carried out a series of experiments with several EDITS configurations (always using the TED algorithm). Two considerations arise from the analysis of the output files. First, as it consists of a purely syntactic representation, some lexical and morphological paraphrases, and especially the semantic ones,[8] do not find an adequate representation. Moreover, paraphrase mechanisms based on pure changes of order are not reflected in the output, as word order is generally not taken into account in dependency analysis.

Second, the tree alignment is, on many occasions, inadequate. In Figure 1 on the left, we see how the 'technologies' node, present in both trees, is not aligned, because it does not occupy the same (or similar) position in the tree.[9] The expected alignment from the paraphrase point of view appears at the right hand representation. Such alignment problems do not have a straightforward solution in the EDITS framework, because they arise from the TED algorithm itself: it is derived from the image recognition literature and it tends to match structure more than content. Once this problem was identified, the next step was to quantify it to evaluate its scope.

3.2 The Tree Alignment Problem

Here we reach our main question of interest: the alignment problem. We compare EDITS with gold standard alignments.

We use Cohn, Callison-Burch, and Lapata (2008)'s paraphrase corpus,[10] as it contains, among other data, 370 positive pairs from the MSRP corpus with manual word or phrase alignments by two annotators (A and C). These annotations constitute the gold standard in our experiments.

In order to be able to carry out the mapping equitatively, we analyzed this same set with EDITS. As the number of pairs is small, we performed 5-fold cross validation.

[8] See Vila, Martí, and Rodríguez (2011) for the paraphrase typology we are referring to. Example (3) above is an example of a semantics based paraphrase.

[9] We understand the nodes connected with the substitution operation as aligned nodes, and the deleted or inserted nodes as non-aligned nodes.

[10] http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_corpus.html

Figure 2 shows an example of gold standard alignment. Horizontal and vertical axes show the sentences of the aligned paraphrase pair. Shadowed squares represent those alignments that the annotators considered Sure (S), in black, and Possible (P), in grey. As can be seen, some of the words remain unaligned ("then" and "Texas" in the vertical sentence) because they do not have a counterpart, and others are aligned in block ("at the FAA" and "lost-aircraft" in the vertical sentence) as there is not a one-to-one word alignment. In Figure 2, we can also see EDITS alignments for the same pair of sentences (E). As can be seen, not all gold standard alignments are covered by EDITS.

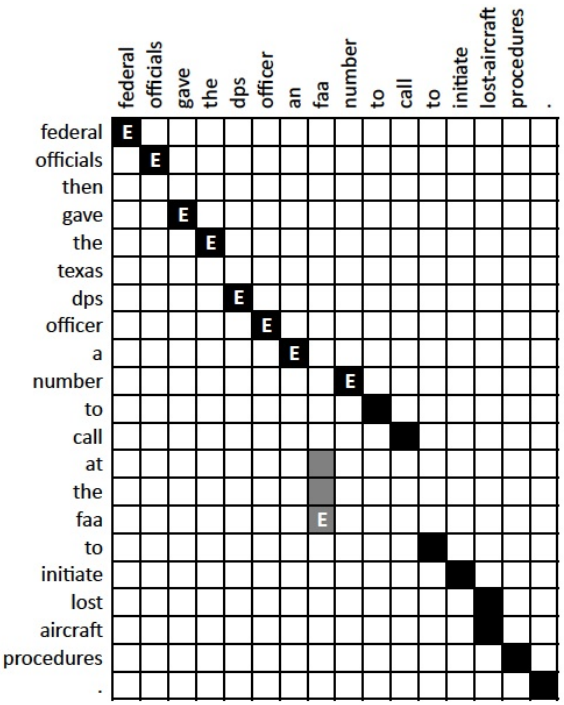

Figure 2: EDITS (E) and gold standard alignments corresponding to annotator C (black, (S)ure; grey (P)ossible) for the sentence paraphrase pair "Federal officials gave the DPS officer an FAA number to call to initiate lost-aircraft procedure" (horizontal axis) and "Federal officials then gave the Texas DPS officer a number to call at the FAA to initiate lost aircraft procedures" (vertical axis).

We performed the mapping between EDITS and gold standard alignments automatically, and computed precision, recall and F1. As can be seen in the first column in Table 1, precision is high (around 0.87) and the recall is low (around 0.50). EDITS only covers a

half of the expected alignments, but the ones that carries out are mainly correct.

We performed a second calculation of the results applying a series of filters in order to get more precise results. We did not consider cases that were erroneously penalized by EDITS in the first calculation. Specifically, we filtered block or discontinuous alignments, which refer to those cases in the gold standard in which a group of (discontinuous) words is aligned to a word or another group of (discontinuous) words. An example of this can be seen in *FAA/at the FAA* and *lost aircraft/lost-aircraft* in Figure 2. This situation cannot take place in EDITS, as the alignment is performed between nodes. We also filtered prepositions, conjunctions and punctuation marks. These elements are aligned in the gold standard, but do not appear as nodes and, thus, are not aligned in our analysis with EDITS. In the case of block alignments, the filter has a linguistic motivation as well: when the gold standard annotators use a block, it is because a word by word alignment is not possible. This case corresponds, on many occasions, to the semantic paraphrases, that cannot be treated with the TED approach (see Section 3.1). As can be seen in the second column in Table 1, once the filters applied, the precision and especially the recall rise (0.1 and more than 0.25 points, respectively).

We also analyzed the cases annotated as S in the gold standard separately. Although the recall rises again, the precision is lower. The reason for this decrease is that some EDITS alignments coincide with P alignments in the gold standard. When we do not take into consideration P alignments, these EDITS alignments are still there, which causes a decrease in the precision.

	− **Filters**		+ **Filters**		+ **Filters Only S**	
	A	*C*	*A*	*C*	*A*	*C*
Precision	0.86	0.88	0.87	0.89	0.85	0.87
Recall	0.50	0.49	0.77	0.78	0.78	0.80
F1	0.63	0.63	0.81	0.83	8.82	0.83

Table 1: EDITS alignment results classified according to the mapping with annotations by annotators A and C in the gold standard.

4 Conclusions

In this paper, we analyzed TED as a baseline approach for paraphrase representation, which may be used as the basis for further work on other approaches to paraphrase representation. As it consists of a purely syntactic approach, paraphrase alternations not based on syntactic reorganizations do no find an adequate representation. Moreover, further work needs to be done in order to improve tree alignments.

We showed that the EDITS suite, initially developped for RTE, can also be applied to the paraphrase task. As a result of the experiments, we obtained the MSRP corpus and a fragment of the Cohn, Callison-Burch, and Lapata (2008)'s corpus processed with EDITS, as well as a mapping between EDITS and Cohn, Callison-Burch, and Lapata (2008) alignments.

5 Future Work

A possible future line of research is the exploration of Tree Alignment Distance (Bille, 2003) and/or (Fanout) Weighted Tree Edit Distance (Augsten, Böhlen, and Gamper, 2010) algorithms, as we hypothesize that they can do better in terms of tree alignment.

Moreover, we plan to work on an approach dealing with paraphrase complexity in a more comprehensive way. Our objective is setting a paraphrasability measure based on the combination of relatedness measures associated to different types of the paraphrase typology by Vila, Martí, and Rodríguez (2011). EDITS would be used to build one of these dimensions.

References

Aho, Alfred V. and Jeffrey D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3(1):37–56.

Augsten, Nikolaus, Michael Böhlen, and Johann Gamper. 2010. The pq-gram distance between ordered labeled trees. *ACM Transactions on Database Systems*, 35(1):art. 4.

Bille, Philip. 2003. Tree edit distance, alignment distance and inclusion. IT University Technical Report Series.

Casacuberta, Francisco and Enrique Vidal. 2007. Learning finite-state models for

machine translation. *Machine Learning*, 66:69–91.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague/Paris.

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

Dolan, William B. and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, pages 9–16.

Dras, Mark. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Australia.

Eisner, Jason. 2003. Learning non-isomorphic tree mappings for machine translation. In *The Companion Volume to the Proceedings of the ACL*, pages 205–208.

Heilman, Michael and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases and answers to questions. In *Proceedings of the HLT-NAACL*, pages 1011–1019.

Kouylekov, Milen and Bernardo Magnini. 2005. Recognizing textual entailment with tree edit distance. In *Proceedings of the PASCAL RTE Challenge*, pages 17–20.

Shasha, Dennis and Kaizhong Zhang. 1990. Fast algorithm for the unit cost editing distance between trees. *Journal of Algorithms*, 11:581–621.

Smith, David and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 23–30.

Vila, Marta, M. Antònia Martí, and Horacio Rodríguez. 2011. Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.

Wan, Stephen. 2010. *Sentence Augmentation: A Text-to-Text Generation Component for Summarisation*. Ph.D. thesis, Macquarie University, Australia.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

Wu, Dekai. 2010. Alignment. In Nitin Indurkhya and Fred Damerau, editors, *Handbook of Natural Language Processing*. Chapman & Hall/CRC, second edition, pages 367–408.

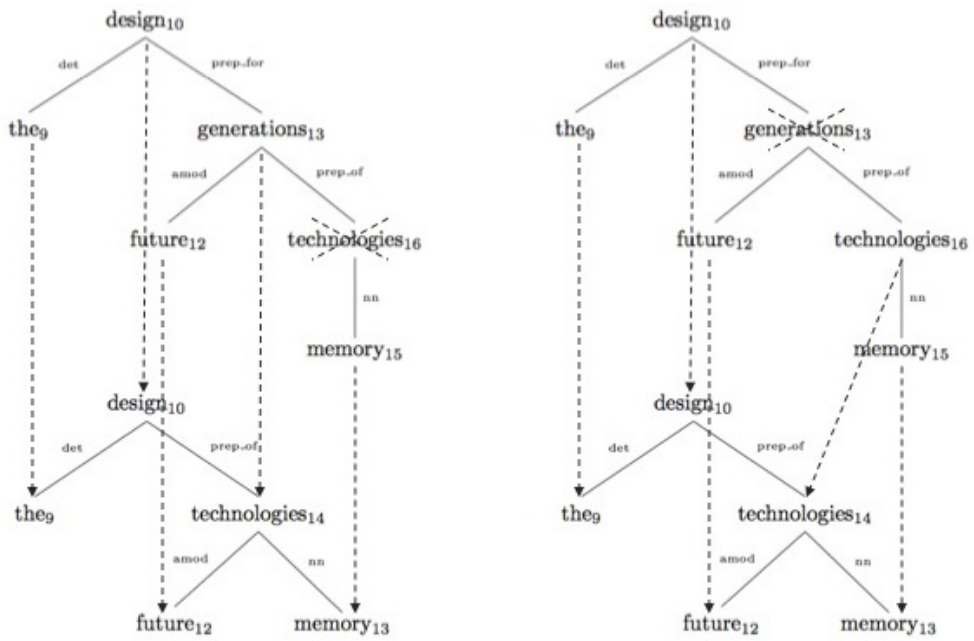

Figure 1: Representation by EDITS (left) and the expected one (right) for the MSRP corpus paraphrase pair (fragment) "the design for future generations of memory technologies" (top) and "the design of future memory technologies" (bottom). Arrow: substitution/alignment; cross: deletion.