An Optimal Quadratic Approach to Monolingual Paraphrase Alignment

Mihai Lintean
Department of Computer Science
The University of Memphis
Memphis, TN 38138, USA
linteanm@yahoo.com

Vasile Rus
Department of Computer Science
The University of Memphis
Memphis, TN 38138, USA
vrus@memphis.edu

Abstract

We model the problem of monolingual textual alignment as a Quadratic Assignment Problem (QAP) which simultaneously maximizes the global lexico-semantic and syntactic similarities of two sentence-level texts. Because QAP is an NP-complete problem, we propose a branch-and-bound approach to efficiently find an optimal solution. When compared with other methods and studies, our results are competitive.

1 Introduction

Textual alignment between two sentences involves the identification of words and phrases considered to be semantically equivalent or very close in meaning (within the context of the respective sentences). Monolingual alignment is particularly useful for the task of text-to-text semantic similarity (Agirre et al., 2012; Rus et al., 2013). Figure 1 shows an example of human generated alignments between two sentences from the corpus used by Thadani et al. (2012), which is a modified corpus of human-aligned paraphrases initially described in Cohn et al. (2008).

While monolingual text alignment has been tackled as a task of its own only recently (MacCartney et al., 2008; Thadani and McKeown, 2011; Yao et al., 2013; Sultan et al., 2014), text alignment has been explored intensely in the area of machine translation (Och and Ney, 2003; Brunning, 2010). Brunning (2010) distinguishes among three levels of alignment in machine translation: document alignment, sentence alignment, and word/phrase level alignment. We focus here on word-level alignment. Furthermore, we focus on monolingual word alignment in the context of sentence-to-sentence similarity tasks such as textual entailment and paraphrase identification.

We focus on word-level (as opposed to phrase-level) alignment for a number of reasons. First, the vast majority of gold alignments in the two datasets we use (95-96%) are word-level alignments (the rest are phrase-level). Similarly, Yao et al. (2013) report that word-level alignments constitute more than 95% of the alignments in recent human-annotated corpora. A second reason is the fact that our formulation of the monolingual alignment task based on the Quadratic Assignment Problem (QAP) (Burkard et al., 1998; Lawler, 1963; Koopmans and Beckmann, 1957) fits well with word-level alignment. Third, the key ingredients in our solution (the word-to-word semantic similarity measures and dependency relations) apply directly to words.

The role of word-to-word semantic similarity measures and contextual information for monolingual alignment has been explored in the past. However, the jury is still out there with respect to how to best combine these types of information for monolingual alignment as one of the most recent work in this area has illustrated (Sultan et al., 2014). Sultan et al. (2014) showed that use of local contextul information in combination with hand-crafted dependency type equivalences yields better results than methods that exploit local context, e.g. Yao et al. (2013). Indeed, our approach combines in unique ways word-to-word semantic similarity measures with contextual information in the form of dependency-relations among words and with a combinatorial optimization formulation based on the QAP problem. As dependencies can capture longer-distance relationships between words in a sentence, we can say that our method uses more than just local context for aligning texts. Furthermore, because the QAP formulation provides a global optimal solution, our method is indeed accounting for the full sentential context.

Indeed, our QAP formulation simultaneously accounts for word-level similarities and similari-

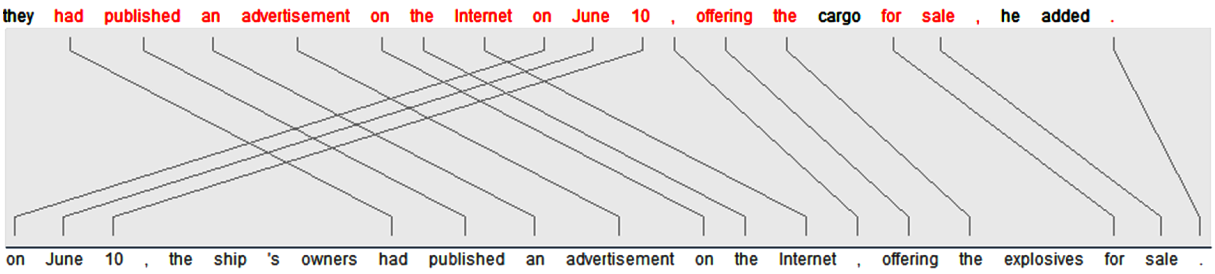

Figure 1: Example of Monolingual Text Alignment (instance #28 of the Edinburg corpus)

ties between corresponding syntactic/grammatical relations in a globally optimal manner. In contrast, Chambers et al. (2007) method for sentence level monolingual alignment finds a local maximum, which only in certain, lucky circumstances may also be a global maximum. Optimization methods have been proposed for phrase-level monolingual alignment (MacCartney et al., 2008; Thadani and McKeown, 2011; Thadani et al., 2012) in the context of a paraphrase task that rely on integer linear programming. Our optimization method is based on a different paradigm, the QAP formulation, and we rely on word-to-word semantic similarity measures, some of which are totally unsupervised such as Latent Semantic Analysis, and syntactic relation identity as opposed to edit distances. Thadani et al. (2011; 2012) used string similarity and WordNet for computing semantic relatedness.

We evaluated the proposed method on two datasets. The first one is the SEMILAR corpus (Rus et al., 2013), a subset of 701 randomly selected pairs from the Microsoft Research Paraphrase Corpus (MSRP) (Dolan et al., 2004). The pairs were manually annotated with tokens and phrase alignments. The second dataset is the evaluation corpus used by Thadani et al. (2012), called the Edinburg corpus, a modified corpus of human-aligned paraphrases, initially described in Cohn et al. (2008).

2 Related Work

The Quadratic Assignment Problem (QAP) is a classical combinatorial optimization problem (Burkard et al., 1998; Lawler, 1963; Koopmans and Beckmann, 1957). QAP has been originally formulated to minimize the overall cost of economic activities. QAP is an NP-hard problem (Sahni and Gonzalez, 1976).

We adapted the QAP formulation to our monolingual sentence-level alignment problem. In our case, we want to find a mapping between words in one sentence to words in another sentence that maximizes the similarity between two texts in terms of word-level similarity and simultaneously accounting for the relations between the matched words. That is, we prefer matchings between words in two texts T_1 and T_2 that not only lead to best word-level similarities but also the dependencies among words in T_1 and the corresponding matched words in T_2 must be optimally accounted for. We use word-to-word similarity measures for quantify the degree to which two words semantically match each other. We experimented with WordNet word-to-word similarity metrics (Pedersen et al., 2004) and the algebraically-derived Latent Semantic Analysis vectorial representation. To extract dependency relations we employed the Stanford CoreNLP Library.

Efforts to optimize the lexico-semantic, i.e. word-level, similarity between texts have been reported. Chan and Ng (2008) proposed a machine translation evaluation metric based on the optimal algorithm for bipartite graph matching also known as the assignment problem (Kuhn, 1955; Munkres, 1957). The assignment problem ignores interdependencies between words in a text although they could be accounted for indirectly, as Chan and Ng did. However, the indirect account of interdependencies among words in a text does not lead to an optimal solution that simultaneously maximizes overall word-level similariy while accounting for their contextual relations as encoded by, for instance, dependency information. QAP has been applied to the problem of word alignment by Lacoste-Julien and colleagues (2006), though their study is applied on pairs of bilingual sentences (i.e. French and English) and it does not consider syntactic dependencies between words in a sentence.

As already mentioned, QAP is an NP-hard problem. Efficient solutions work in general up to problem sizes of 25 using dynamic programming or branch-and-bound methods. In our case, we propose a branch-and-bound method which guarantees the finding of optimal solutions for short texts, i.e. typical sentences as those found in the SEMILAR and Edinburg corpora, our target datasets.

3 Alignment Approaches

We present in this section the details of the proposed optimal solution to the task of textual alignment in the context of semantic similarity of two sentences based on the QAP formulation. We start by describing two simpler solutions for monolingual text alignment: a greedy approach and an optimal solution based on the assignment problem for which a polynomial algorithm exists – the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957). We evaluated and compared these simpler alignment methods to the proposed QAP solution. In our experiments, we followed a train-test methodology in which we first learnt the parameters of the various approaches on the training part of the data sets and then used the trained approaches to evaluate the QAP method on the test portion of the data.

3.1 Greedy Word-to-Word Alignment (GRD)

In the greedy approach to monolingual alignment, words from one sentence (usually the shorter sentence) are greedily matched, one by one, starting from the beginning of the sentence, with the most similar word from the other sentence. In case of duplicates, because we require that words must be part of at most one pair the order of the duplicate words in the two sentences becomes important such that the first occurrence in one sentence matches with the first occurrence in the other sentence and so on. Otherwise, the order in which the matching words appear in the two sentences does not matter. While simple and fast, the obvious drawback of the greedy method is that it can mistakenly match words if there are two or more ways to pair them, simply because of the order in which they were processed. The next method tries to solve this problem by searching for an alignment that leads to a global maximum similarity score across all pairs of aligned words.

3.2 Optimal Word-to-Word Alignment via Assignment Problem (w-OPT)

The job assignment problem or sailor assignment problem or just the assignment problem is one of the fundamental combinatorial optimization problems and consists of finding a maximum weight matching in a weighted bipartite graph. Given a complete bipartite graph, $G = (S, T, E)$, with n sailor vertices (S), n ships vertices (T), and each edge $e_{s \in S, t \in T} \in E$ has a non-negative weight $w(s, t)$ indicating how qualified a sailor is for a certain job, the task is to find a matching M from S to T with maximum weight. In case of different numbers of sailors or ships, dummy vertices could be used.

The assignment problem can be thus formulated as finding a permutation π for which $S_{w-OPT} = \sum_{i=1}^{n} w(s_i, t_\pi(i))$ is maximum. Such an assignment is called optimum assignment. An algorithm, the Kuhn-Munkres method (Kuhn, 1955), has been proposed that can find a solution to in polynomial time.

In our case, we model the semantic similarity problem as finding the optimum assignment between words in one text, T_1, and words in another text, T_2, where the fitness between words belonging in opposite texts can be measured by any word-to-word semantic similarity function. That is, we are after a permutation π for which $S_{w-OPT} = \sum_{i=1}^{n} \Theta_{sim}(v_i, w_{\pi(i)})$ is maximum where we note Θ_{sim} to be any word-to-word similarity measure, and v and w are words from the texts T_1 and T_2, respectively.

The assignment problem only focuses on optimally matching words in one sentence S to words in the other sentence T based only on how the words in S match the words in T. Interdependencies among words in S or among words in T are not taken into account. A solution that simultaneously accounts for such inter-dependencies, thus capturing the context of each word in their corresponding sentences, is presented next.

3.3 Optimal Sentence Alignment via Quadratic Assignment (QAP)

QAP has two well-known, historically important, formulations: the Koopmans-Beckmann (1957) formulation, and the more general Lawler (1963) formulation. We adapted the Koopmans-Beckmann (1957) formulation as it more clearly fits our task.

The goal is to find the optimum placement function π that maximizes the objective function $QAP(F, D, B)$ defined below where F and D describe syntactic dependencies between words in one sentence (S) and the other (T), respectively, while B captures the word-to-word similarity between words across the two sentences, all of them being symmetric, non-negative matrices. It should be noted that in the original formulation the objective function was about minimizing an economic cost while in our formulation we maximize the semantic similarity between two sentences. We further extend the objective function QAP by adding relative weighs to both terms in the above formulation resulting in the formulation below:

$$\max QAP_{(F,D,B)} = \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} f_{i,j} d_{\pi(i)\pi(j)} + \\ +(1-\alpha) \sum_{i=1}^{n} b_{i,\pi(i)}$$

The $f_{i,j}$ term quantifies the syntactic relation between words i and j in text T_1 which are mapped to words $\pi(i)$ and $\pi(j)$ in text T_2, respectively. The distance $d_{\pi(i)\pi(j)}$ quantifies the syntactic relation between words π_i and π_j. For words i and j that have a direct dependency relation, $f_{i,j}$ is set to 1 and 0 in case there is no direct dependency between the two words. Similarly, the distance $d_{\pi(i)\pi(j)}$ between words $\pi(i)$ and $\pi(j)$ is set to 1 in case there is a direct dependency relation among them and 0 otherwise. We also experimented with a variant in which we enforced that the dependency between words i and j and the dependency between the corresponding words in the other text, $\pi(i)$ and $\pi(j)$, be of same type. That is, we prefer matchings between words in two texts T_1 and T_2 that not only lead to direct dependencies between words in T_1 and direct dependencies between corresponding matched words in T_2 but those dependencies must be of the same type. We obtained best results with this latter version which we used to generate all results in this paper.

The α parameter can be used to bias the search, to look for solutions that give more weight to matching dependencies (represented in the first term of the objective function) than to word similarities (represented in the second term), or vice-versa. When $\alpha = 0.5$, equal importance is given to both alignment criteria.

Solution Space

A brute force solution to the QAP problem, which would generate all possible mappings from words in a sentence to words in the other sentence, i.e. all permutations, is infeasible as the solution space is too big. When considering all possible pairings of words between sentence A, of size n, and sentence B of size m, where $n < m$, and we pose no limitations on the type of pairings that can be made, there are $m!/(m-n)!$ possible solutions. It should be noted that exact proposed solutions to the QAP problem can only handle instances up to $n = 25$ (Christofides and Benavent, 1989) or in special cases up to $n = 30$ (Anstreicher, 2003).

In the case of sentences of average size n=m=20 words, there are $2.4{*}10^{18}$ possible pairings, which is too large. We have taken a number of steps to reduce the solution space in our case. We know that words can only be paired with other words that are semantically similar. Given a word-to-word similarity metric, Θ_{sim}, which outputs a normalized similarity value between 0 and 1 (0 means not similar, 1 indicates equivalent meaning), we can impose to pair only words with Θ_{sim} greater than a similarity threshold value, which we will denote Ω. For instance, it does not make sense to consider matching a verb with a determiner even if the similarity is non-zero (but very close to zero, e.g. LSA similarity score between *provide* and *the* is 0.042). Moreover, in regard to the initial QAP search, we can further reduce the space by focusing on pairing only numbers and content words (i.e. nouns, verbs, adjectives and adverbs), as these, along with their associated dependencies, carry most of the relevant semantic content in a sentence. We also chose to pair words that have either identical lemma forms (i.e. *tell* vs. *told*, *gains* vs. *gain*) or the same part of speech. These constraints reduce considerably the QAP search space.

The average size of a sentence in the SEMILAR corpus is 21 tokens, with a maximum of 38, while the Edinburg corpus contains sentences with an average length of 22 tokens and a maximum of 50. Our branch and bound search allowed us to find an optimal solution in under a second for all instances on both corpora, when $\alpha \leq 0.5$.

Solving QAP via Branch & Bound

Branch-and-bound is one of the widely used paradigms for handling NP-hard optimization problems such as QAP. The gist of the branch-and-bound paradigm is to avoid explicitly exploring

the entire solution space, which is too big for NP-hard optimization problems, while assuring that the unexplored parts of the space cannot contain the optimal solution. This is possible by defining a bounding function that always overestimates or underestimates solutions, depending on what type of optimal solution is sought, maximum or minimum cost, respectively.

The proposed branch-and-bound method starts with an initial solution, e.g. the optimal word-to-word matching approach obtained using the Kuhn-Munkres algorithm (w-OPT). We call this the current optimal QAP solution (C; optimal solution so far) and we denote C_{QAP} the value of the $QAP(F, D, B)$ objective function for this solution. Next, the method iteratively explores new solutions comparing at each step the current optimal solution with new ones. The exploration follows a search tree where each node represents a subspace of solutions. In our case, a subspace is defined by a partial pairing, P, with p word-to-word assignments ($p < n$). We define a bounding function $F(P)$ to compute an upper bound for the partial pairing and therefore for the entire subspace of solutions that contains this partial pairing. That is, any solution S containing the partial pairing P will have a QAP score that is guaranteed to be less than the value of the bounding function for the current node in the search tree, $S_{QAP} \leq F(P)$, for $\forall S, S \supseteq P$. If $F(P)$ is not greater than the best solution found so far, i.e. $F(P) \leq C_{QAP}$, it means there is no better solution than C_{QAP} within the subspace of complete solutions that contain the partial pairing P, as $F(P)$ always overestimates the QAP score of the solutions in this subspace. Thus, the entire subspace can be further ignored from the search. The details of the bounding function $F(P)$ are not presented here due to space reasons.

Comparing QAP Alignment with GRD and w-OPT

In this subsection, we exemplify how quadratic assignment (QAP) is more powerful when it comes to aligning words in two sentences than the other two methods described earlier: greedy (GRD) and optimal word matching (w-OPT). We take the example previously shown in Figure 1, an actual sample instance extracted from the Edinburg corpus. Its greedy alignment is shown in Figure 2. Because of the selective order in which the greedy method picks the matchings, notice that the two

Θ_{sim}	Ω	Method	Prec	Recall	F1
LCH	0.8	GRD	93.45	84.18	88.04
		w-OPT	93.94	84.62	88.51
		QAP	95.39	86.44	90.17
LSA	0.4	GRD	93.42	84.01	87.95
		w-OPT	94.07	84.56	88.55
		QAP	95.55	86.14	90.10
JCN	0.1	QAP	90.78	87.82	88.75
	0.2		94.70	87.15	**90.26**
	0.4		95.24	86.62	90.21
	0.6		95.39	86.45	90.18
	0.8		95.37	86.39	90.14
	0.9		95.45	86.38	90.17
		METEOR	94.22	84.77	**88.64**

Table 1: Alignment percent scores on SEMILAR corpus

'*the*' determiners, the '*on*' prepositions and both commas are mistakenly matched between the two texts. The word optimal (w-OPT) method does not perform any better in this case. The QAP method however, through the right use of the syntactic dependencies, is almost identical with the human annotations shown in Figure 1, except that it finds one extra unneeded pair between commas.

Though for our example, the w-OPT method does not perform any different than the greedy method, from our experiments we found that it does perform better overall, but not consistently better. This is due to the high-lexical overlap between sentences to be aligned in the datasets we used.

4 Experiments and Results

We evaluated and compared the three alignment methods presented in the previous section (GRD, w-OPT and QAP) on two datasets that were manually annotated with alignments between sentences: the SEMILAR corpus (Rus et al. 2013) and the Edinburg corpus (Thadani et al. 2012). The SEMILAR corpus consists of a set of 701 instances extracted from the MSRP corpus and which were tokenized, tagged and parsed with the Stanford Core NLP library and then manually annotated with tokens and phrase alignments. The Edinburg corpus contains 714 annotated instances used for training, and 306 instances used for evaluation, also pre-processed and parsed for syntactic dependencies using the Stanford NLP Parser.

As in Thadani et al. (2012), we used

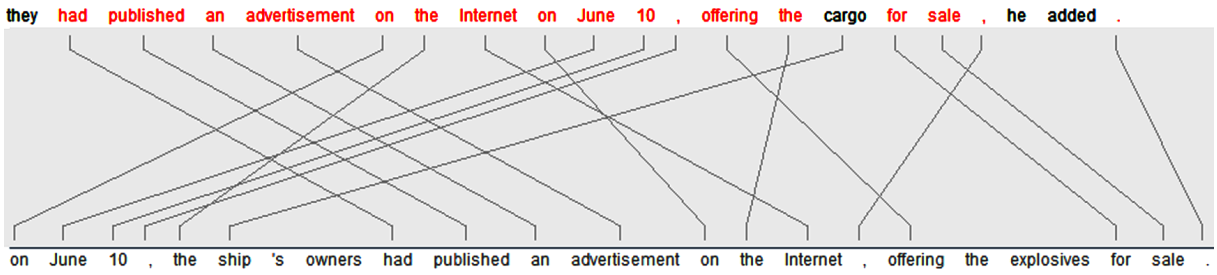

Figure 2: Greedy alignment on training instance #28 of the Edinburg corpus

METEOR's maximum accuracy alignment (Denkowski and Lavie, 2011) as a baseline to compare with our alignments. The evaluation scores for the alignments are also similarly computed as macro-average results: precision, recall, and F-score values are computed for each instance, then these scores are averaged across all instances. Because we do word-level alignments and the human-annotated data and METEOR output include phrase-level alignments we have to have a way to consistently assess the output of the method. We assessed the phrase-level alignments using word-level alignments as explained next. If the gold data contains a phrase-level alignment then if the output of a method contains an alignment between any two words in the gold-aligned phrases then we consider the system word-level alignment as a hit. Using this method, the METEOR alignments are evaluated at word-level and therefore can be directly compared to our methods' alignments. It should be noted that phrase-level alignments are very few. On the Edinburg corpus there are only 95 phrase alignments produced by METEOR out of 5,046 alignments (word- and phrase-level) and on the SEMILAR corpus METEOR produces only 30 phrase-level alignments out of 10,112 alignments. This method of evaluating neither penalizes nor rewards METEOR.

4.1 Results on the SEMILAR Corpus

Table 1 shows the alignment performance results on the SEMILAR corpus, for all three alignment methods and the METEOR baseline. For space reasons, we picked two representative word-to-word similarity metrics, JCN (Pedersen et al., 2004) and LSA, and report comparative results among the three alignment methods. Also, we illustrate the impact of the Ω parameters using the QAP method and a third word-to-word metric, JCN (Pedersen et al., 2004). Note that by changing the Ω value within some restrictive bounds, one could control for a better precision, at the expense of the recall, or viceversa, while keeping the overall F-score more or less the same. The other word-to-word metrics that we experimented with, show a similar trend in performance, with very small variations from the ones we reported. It is important to note that for the QAP method we used $\alpha = 0.5$ which was chosen following the same process explained in the next section. The QAP method significantly outperforms both GRD and w-OPT alignments for both JCN and LSA word-to-word similarity metrics ($p < 0.0018$). The difference in performance between GRD and w-OPT is significant only on the LSA metric ($p < 0.0058$). Note that the high performance scores for all methods are due to the high lexical overlap, a characteristic of the SEMILAR instances, which was inherited from the original MSRP corpus.

4.2 Results on the Edinburg Corpus

We present now results when evaluating the three alignment methods on the Edinburg corpus. As a first step, we used the optimal method (w-OPT) on the training subset to find the optimal word-to-word threshold (Ω) for seven word similarity (Θ_{sim}) metrics. Six of them are WordNet based: LIN, PATH, JCN, LCH, RES and WUP (Pedersen et al., 2004); and one is LSA. Word-to-word threshold values between 0 to 1 were evaluated in increments of 0.01 and the ones that gave the best F-Score on the training set, when using the w-OPT method, were selected: $\Omega(LIN) = 0.73$, $\Omega(PATH) = 0.3$, $\Omega(JCN) = 0.23$, $\Omega(LCH) = 0.69$, $\Omega(RES) = 0.47$, $\Omega(WUP) = 0.85$, $\Omega(LSA) = 0.1$.

Next, we searched for a good parameter α value to use in the QAP alignment. We evaluated QAP on the training set for several α values, from 0 to

Θ_{sim}	Ω	Method	Prec	Recall	F1
LIN	0.73	GRD	88.06	78.64	82.51
		w-OPT	89.18	79.14	83.30
		QAP	90.95	84.15	**86.87**
JCN	0.23	GRD	88.17	78.52	82.47
		w-OPT	89.33	79.02	83.27
		QAP	90.92	83.9	86.70
PATH	0.30	QAP	89.42	84.86	86.51
LCH	0.69		90.39	84.39	86.73
RES	0.47		92.82	80.57	85.61
WUP	0.85		90.75	84.19	**86.78**
LSA	0.10		88.94	84.63	86.24
		METEOR	88.10	83.37	**85.22**

Table 2: Alignments percent scores on Edinburg corpus

1 in increments of 0.1, and various word-to-word metrics. We found that $\alpha = 0.5$, which gives equal importance to both word and dependency relations, is the optimal value that maximizes F-measure on training and therefore we used this value for the test data.

Finally, we evaluated all three alignments methods on the testing part of the Edinburg corpus. We ran paired t-tests on the alignment performances (w-OPT against GRD, and QAP against w-OPT). We found QAP results to be statistically significantly better than w-OPT ($p < 0.0001$), and w-OPT to be significantly statistically better than GRD ($0.005 > p > 0.0004$) across all the seven word metrics that we used.

We also ran t-tests between the results given by our best word metric, LIN, and the other metrics. We found the differences were not statistically different, except on the LSA metric ($p = 0.0281$), and RES ($p < 0.0001$).

Table 2 shows comparative performance results for all three alignment methods on only two word metrics, LIN and JCN, for space reasons, and QAP comparative results for all the other word metrics, along with the METEOR alignment results.

It should be noted that the results reported by Thadani et al. (2012) consider phrase-level alignment and therefore their results are not directly comparable to ours. They report results slightly worse than METEOR on precision (-5%) and considerably better on recall ($+10\%$). For our case, we found that the QAP method is consistently better than METEOR, in terms of all measures, on all the word metrics except RES, which although gives best precision, it is highly penalized on recall.

5 Discussion and Conclusions

We proposed in this paper a novel approach to the task of aligning monolingual texts in the context of semantic similarity tasks based on an efficient branch and bound approach. We showed that our optimal solution provides state-of-the-art performance. Although the proposed method is computationally more expensive, it consistently outperforms other alignment methods and provides optimal solutions for sentences of average size (< 40 words). The proposed QAP solution can be useful for a number of tasks such as semantic similarity assessment and phrase level semantic equivalence extraction and in many applications such as intelligent tutoring systems, question answering, and automated essay scoring.

Acknowledgments

This research was supported in part by the Institute for Education Sciences under award R305A100875. Any opinions, findings, and conclusions or recommendations are solely the authors'.

References

Eneko Agirre, Daniel Cel, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. *Proceedings of SemEval 2012, in conjunction with *SEM 2012.* Montreal, Canada, June 7-8, 2012.

Kurt M. Anstreicher. 2003. Recent advances in the solution of quadratic assignment problems. *Mathematical Programming*, volume 97(1-2):27–42. Springer.

James Brunning. 2010. *Alignment Models and Algorithms for Statistical Machine Translation*. Ph.D. Thesis, Cambridge University Engineering Department.

Rainer E. Burkard, Eranda Çela, Panos M. Pardalos, and Leonidas S. Pitsoulis. 1998. The Quadratic Assignment Problem. In P.M. Pardalos and D.Z. Zu, editors, *Handbook of Combinatorial Optimization*, volume 3:241–337. Kluwer Academic Publishers.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine De Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning Alignments and Leveragning Natural Logic. In *Proceedings of the ACL-07 Workshop on Textual Entailment and Paraphrasing*.

Yee S. Chan and Hwee T. Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62.

Nicos Christofides and Enrique Benavent. 1989. An exact algorithm for the quadratic assignment problem. *Operations Research*, volume 37(5):760–768.

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

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3 Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *EMNLP 2011 - Workshop on Statistical Machine Translation*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland.

Tjalling C. Koopmans and Martin Beckmann. 1957. Assignment Problems and the Location of Economic Activities. *Econometrica*, volume 25(1):53–76.

Harold W. Kuhn. 1955. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, volume 2:83–97.

Simon Lacoste-Julien, Ben Taskbar, Dan Klein, and Michael I. Jordan. 2006. Word Alignment via Quadratic Assignment. In *Proceedings of HLT-NAACL '06*, pages 112–119. New York, June 2006.

Eugene L. Lawler. 1963. The quadratic assignment problem. *Management Science*, volume 9:586–599.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A Phrase-Based Alignment Model for Natural Language Inference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu. October, 2008.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38. Society for Industrial and Applied Mathematics.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, volume 29:19–51.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *NAACL-2004*.

Vasile Rus, Mihai Lintean, Rajendra Banjade, Nobal Niraula, and Dan Stefanescu. 2013. SEMILAR: The Semantic Similarity Toolkit. In *Proceedings of ACL 2013*. Sofia, Bulgaria. August 4-9, 2013.

Sartaj Sahni and Teofilo Gonzalez. 1976. P-complete approximation problems. *Journal of the Association for Computing Machinery*, volume 23:555–565.

Arafat Md Sultan, Steven Bethard, and Tamara Summer 2014. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the Association of Computational Linguistics*, volume 2(1):219–230.

Kapil Thadani, Scott Martin, and Michael White. 2012. A Join Phrasal and Dependency Model for Paraphrase Alignment. In *Proceedings of COLING-2012*

Kapil Thadani and Kathleen McKeown. 2011. Optimal and syntactically-informed decoding for monolingual phrase-based alignment. In *Proceedings of ACL-HLT, HLT '11*, pages 254–259.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch and Peter Clark 2013. A Lightweight and High Performance Monolingual Word Aligner. In *ACL (2)*. The Association for Computer Linguistics, pages 702–707.