# A Joint Phrasal and Dependency Model for Paraphrase Alignment

*Kapil THADANI*[1] *Scott MARTIN*[2] *Michael WHITE*[2]
(1) COLUMBIA UNIVERSITY, New York NY 10027
(2) OHIO STATE UNIVERSITY, Columbus OH 43210
kapil@cs.columbia.edu, {scott,mwhite}@ling.ohio-state.edu

ABSTRACT
Monolingual alignment is frequently required for natural language tasks that involve similar or comparable sentences. We present a new model for monolingual alignment in which the score of an alignment decomposes over both the set of aligned phrases as well as a set of aligned dependency arcs. Optimal alignments under this scoring function are decoded using integer linear programming while model parameters are learned using standard structured prediction approaches. We evaluate our joint aligner on the Edinburgh paraphrase corpus and show significant gains over a Meteor baseline and a state-of-the-art phrase-based aligner.

TITLE AND ABSTRACT IN FRENCH

## Un modèle de phrases et de dépendances pour l'alignement des paraphrases

L'alignement monolingue s'impose fréquemment dans les tâches de langue naturelle qui comprennent des phrases similaires. Nous présentons un nouveau modèle pour l'alignement monolingue dans lequel le score d'un alignement tient compte de l'ensemble de phrases alignées et d'un ensemble d'arcs de dépendance alignés. Cette fonction de score donne des alignements en utilisant l'optimisation linéaire, et nous effectuons l'apprentissage des paramètres du modèle avec des méthodes standardes de prédiction structurée. Nous évaluons notre système mixte par rapport au corpus de paraphrases d'Edinburgh et nous démonstron un avantage significatif par rapport á Meteor et á un système de pointe fondé sur l'alignement des phrases.

KEYWORDS: monolingual alignment, integer linear programming, structured prediction.

KEYWORDS IN FRENCH: alignement monolingue, optimisation linéaire, prediction structurée.

*Proceedings of COLING 2012: Posters*, pages 1229–1238,
COLING 2012, Mumbai, December 2012.

# 1 Introduction

Textual alignment involves the identification of links between words or phrases which are effectively semantically equivalent in their respective input sentences. *Monolingual* alignment in particular is often needed in natural language problems which involve pairs or groups of related sentences such as textual entailment recognition, multidocument summarization, text-to-text generation and the evaluation of machine translation systems. For example, paraphrase recognition systems can use alignments between input sentences to identify mentions of repeated concepts and determine the degree to which the input sentences overlap.

Recent work on monolingual alignment problems (MacCartney et al., 2008; Thadani and McKeown, 2011) has focused on phrase-based techniques in which the alignment between a pair of sentences is represented through a set of aligned phrase pairs; this has demonstrated advantages over token-based aligners such as Chambers et al. (2007) as well as standard aligners used in machine translation (Och and Ney, 2003; Liang et al., 2006). This paper presents an improved model for monolingual phrase-based alignment that elegantly accounts for syntactic relationships between tokens by additionally considering an *arc-based* alignment representation comprising a set of aligned pairs of dependency arcs consistent with the phrase-based representation. Under this formulation, the score of any alignment is simply defined to factor over all aligned phrase pairs and arc pairs in the alignment. However, recovering a full sentence alignment that optimizes this joint scoring function is non-trivial due to both the interdependence among individual phrase alignments as well as the interaction between phrase-based and arc-based alignments to ensure consistency between the two representations.

In this paper, we describe a technique to recover joint phrasal and arc-based alignments by using integer linear programming (ILP). Given a feature-based scoring function, standard structured prediction techniques can be leveraged to learn parameters that weight features over phrasal and arc-based alignments. We evaluate this joint aligner on a human-annotated paraphrase corpus (Cohn et al., 2008) and show significant gains over phrase-based alignments generated by the Meteor metric for machine translation (Denkowski and Lavie, 2011) as well as a state-of-the-art discriminatively-trained phrase-based aligner (Thadani and McKeown, 2011).

# 2 Related Work

Text alignment is a crucial component of machine translation (MT) systems (Vogel et al., 1996; Och and Ney, 2003; Liang et al., 2006; DeNero and Klein, 2008); however, the general goal of multilingual aligners is the production of wide-coverage phrase tables for translation. In contrast, monolingual alignment is often consumed directly in applications like paraphrasing and textual entailment recognition; this task therefore involves substantially different challenges and tradeoffs.[1] Nevertheless, modern MT evaluation metrics have recently been found to be remarkably effective for tasks requiring monolingual alignments (Bouamor et al., 2011; Madnani et al., 2012; Heilman and Madnani, 2012)—even used off-the-shelf with their default parameter settings—and for this reason we use Meteor as a baseline in this paper.

Monolingual token-based alignment has been used for many natural language processing applications such as paraphrase generation (Barzilay and Lee, 2003; Quirk et al., 2004). Dependency arc-based alignment has seen similar widespread use in applications such as sentence fusion (Barzilay and McKeown, 2005; Marsi and Krahmer, 2005), redundancy removal (Thadani and McKeown, 2008) and textual entailment recognition (Dagan et al., 2005). Furthermore,

[1] See MacCartney et al. (2008) for an enumeration of these challenges in the context of entailment recognition.

joint aligners that simultaneously account for the similarity of tokens and dependency arcs have also been explored (Chambers et al., 2007; Chang et al., 2010). Monolingual phrase-based alignment was first tackled by the MANLI system of MacCartney et al. (2008) and was subsequently expanded upon by Thadani and McKeown (2011) to incorporate exact inference.

ILP has seen widespread use in natural language problems involving formulations which cannot be decoded efficiently with dynamic programming but can be expressed as relatively compact linear programs. DeNero and Klein (2008) and Thadani and McKeown (2011) proposed ILP approaches to finding phrase-based alignments in a multilingual and monolingual context respectively. Chang et al. (2010) describe a joint token-based and arc-based alignment technique using ILP to ensure consistency between the two alignment representations. Our proposed joint phrasal and arc-based aligner generalizes over both these alignment techniques.

## 3 Corpus

As our dataset, we use a modified version of the human-aligned corpus of paraphrases described by Cohn et al. (2008), which we call the *Edinburgh corpus*. We derive this dataset from the original corpus first by standardizing the treatment of quotes (both single and double) and by truecasing the text (Lita et al., 2003). Following MacCartney et al. (2006), we collapse named entities using the Stanford named entity recognizer[2] trained on the pre-built models distributed with it (Finkel et al., 2005). For example, the corpus contains a sentence with the named entity *Bank of Holland*, which we collapse to the single token *Bank_of_Holland*. In future work, we plan to leave the original corpus uncollapsed and annotate named entities by token index.

Our training/testing splits are as follows. We use all of the nonoverlapping portions of the Edinburgh corpus (those only aligned by a single human annotator) as training data. We then randomly sample training instances from the overlapping portions of the corpus: 45 instances from the 'trial' portion drawn from the 'mtc' subcorpus, 19 from the 'news' portion, and 10 from the 'novels' portion. The testing data includes all of the instances in the overlapping portions of the corpus that are not selected as training data, plus the five remaining 'trial' instances. The resulting splits yield 70% for training and 30% for testing, with identical ratios from the three subcorpora ('mtc', 'news', and 'novels') in both training and testing. The training set has 715 paraphrase pairs with a total of 29,827 tokens and an average of 20.9 tokens per sentence, while the test set has 305 paraphrase pairs with 14,391 tokens and 23.6 tokens/sentence on average. Finally, rather than using the merged alignments from the Edinburgh corpus for the overlapping portions, we randomly select one of the two annotators to use as the reference alignment in an unbiased way, with each annotator chosen exactly half of the time.[3]

## 4 Corpus Analysis and Example

Figure 1 shows an example paraphrase pair from the training portion of the corpus. At the top are the Meteor alignments as visualized by the Meteor X-ray tool using shaded boxes, along with the gold standard alignments using filled circles for SURE alignments and open circles for POSSIBLE alignments. Below the alignment grid, the recall errors (SURE only) in the Meteor alignments that are supported by Stanford parser dependencies are shown in bold. These recall errors are supported in the sense that the missed aligned tokens participate in dependencies with other aligned tokens. For example, Meteor fails to align *scout* with *monitor*. This token-level alignment is supported by two aligned dependencies, namely the alignment of

[2] `http://nlp.stanford.edu/software/CRF-NER.shtml`

[3] The modified corpus is available at `http://www.ling.ohio-state.edu/~mwhite/data/coling12/`.

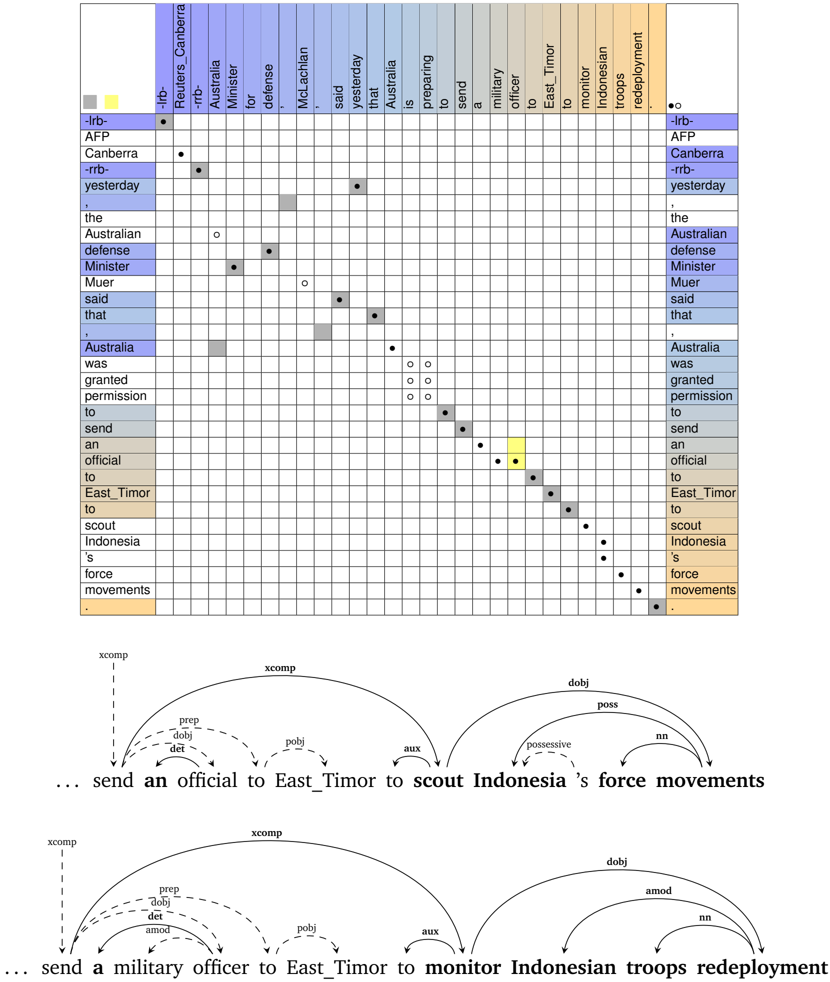


Figure 1: At top, example Meteor alignments (shaded boxes, gray for exact matches and yellow for stem/synonym/paraphrase matches) along with gold SURE and POSSIBLE alignments (circles, filled for SURE and open for POSSIBLE); at bottom, Meteor recall errors (SURE only, in bold) that are supported by aligned Stanford parser dependencies (solid lines).

*send* $\xrightarrow{xcomp}$ *scout* with *send* $\xrightarrow{xcomp}$ *monitor* and *scout* $\xrightarrow{aux}$ *to* with *monitor* $\xrightarrow{aux}$ *to*. Here, the other tokens in the dependencies are identical, and thus the dependencies provide strong evidence for the token-level alignment. Interestingly, the final three recall errors involve interrelated dependencies, suggesting the need for joint inference.

Using this notion of dependency arc alignments supporting token-level alignments, we counted how frequently the token alignments were supported by dependency alignments, and found that 64% of the SURE alignments and 65% of the SURE+POSSIBLE alignments in the training corpus were supported in this way. We also tabulated how often the dependencies were aligned, and found that 54% of the dependency arcs were aligned based on the SURE token alignments, and 62% were aligned based on the SURE+POSSIBLE alignments, thus indicating the greater potential of dependencies to aid alignment when including the POSSIBLEs. The alignment percentages varied considerably by type: of the non-rare dependency types, 74% of the *aux* dependencies were aligned (including the POSSIBLES), while only 38% of the *rcmod* dependencies were aligned, with most core dependency types such as *xcomp* and *dobj* in the 64-70% range.[4]

## 5 Joint alignment framework

Consider a pair of text segments $\langle T_1, T_2 \rangle$ where each $T_s$ represents a set of $n_s$ tokens. We denote $T_s \triangleq \{t_i^s : 1 \leq i \leq n_s\}$ where each $t_i^s$ represents a token in the $i$th position of segment $s$. We also use the notation $t_{i\ldots j}^s \triangleq \{t_k : t_k \in T_s, i \leq k \leq j\}$ to indicate the subsequence of contiguous tokens from positions $i$ to $j$ (inclusive) in $T_s$. Each $T_s$ is also associated with a dependency graph $D_s$ which is treated as a set of labeled arcs, i.e., $D_s \triangleq \{d_{ij}^s : t_j^s \text{ is a dependent of } t_i^s \in T_s \cup \{\textsc{root}\}\}$.

### 5.1 Alignment representations

Our proposed alignment formulation has its roots in the phrase-based representation proposed in MacCartney et al. (2008) and Thadani and McKeown (2011). An alignment $E$ between $T_1$ and $T_2$ is represented by a set of edits $\{e_1, e_2, \ldots\}$ which indicate the modifications that would be needed to convert $T_1$ to $T_2$. We consider two types of edits:

1. *Phrase edits* capture the changes that would need to be made to subsequences of tokens to transform $T_1$ to $T_2$ and vice versa. These are of two types: the first represents the *alignment* of equivalent phrases in $T_1$ and $T_2$ while the other denotes *deletion* or non-alignment of phrases from either $T_s$. A valid phrase-based alignment configuration, denoted by $E_{\text{phr}}$ must have every token participating in exactly one edit.

2. *Arc edits* similarly capture the alignments or deletions of edges in a dependency graph. For a dependency alignment configuration $E_{\text{arc}}$ to be meaningful, the edits in it must be kept *consistent* with the phrase-based alignment configuration $E_{\text{phr}}$. Specifically, two edges that have both their source and target tokens aligned (i.e., participating in the same alignment edit) must also participate in an alignment edit.

We assume that the score for an alignment $E$ factors over the phrase and arc edits present in $E$. Using $e^*$ to represent alignment edits and $e^-$ to represent deletion edits, this can be written as:

$$score(E) = \sum_{e^*_{\text{phr}} \in E} \alpha_{\text{phr}}(e^*_{\text{phr}}) + \sum_{e^-_{\text{phr}} \in E} \delta_{\text{phr}}(e^-_{\text{phr}}) + \sum_{e^*_{\text{arc}} \in E} \alpha_{\text{arc}}(e^*_{\text{arc}}) + \sum_{e^-_{\text{arc}} \in E} \delta_{\text{arc}}(e^-_{\text{arc}}) \tag{1}$$

[4] Note that dependencies can fail to be aligned for a variety of reasons, including parse errors, head-dependent inversions (not taken into account in this paper) and more large-scale structural divergences.

where scoring functions $\alpha_{\text{phr}} : \langle t^1_{i\ldots j}, t^2_{k\ldots l} \rangle \rightarrow \mathbb{R}$ indicate the score of aligning a pair of token sequences, and $\delta_{\text{phr}} : t^s_{i\ldots j} \rightarrow \mathbb{R}$ indicate the score of deleting any token sequence of segment $s$ from the alignment. $\alpha_{\text{arc}} : \langle d^1_{ij}, d^2_{kl} \rangle \rightarrow \mathbb{R}$ and $\delta_{\text{arc}} : d^s_{ij} \rightarrow \mathbb{R}$ are defined analogously for scoring alignments and deletions of arc edits respectively.

## 5.2 Features and learning

The scoring function described above is parameterized by features over the different categories of edits, i.e., $score(E) = \sum_{e \in E} \mathbf{w} \cdot \Phi(e)$ where $\Phi(e)$ is a feature vector for edit $e$ and $\mathbf{w}$ is a vector of parameter weights. The features defined over phrase edits are similar to MacCartney et al. (2008); these encode the type of edit (alignment or deletion), the size of the phrases in alignment edits, the similarity of the phrases determined by leveraging various lexical resources, as well as contextual and positional features. Features for arc edits simply encode the type of edit for an arc of a given class of dependency label, e.g., whether an alignment edit involves two *subj* dependencies, or whether a deletion edit involves a *det* dependency.

Given a inference technique for alignments under the parameterized scoring function, feature weights $\mathbf{w}$ can be learned using any appropriate structured prediction technique. We employ the structured perceptron (Collins, 2002) in our experiments.

## 5.3 Inference via ILP

We now describe an integer linear program that recovers optimal solutions to the problem of jointly recovering a phrasal and arc alignment given any parameter configuration $\mathbf{w}$. Although ILPs in general do not have guarantees on returning solutions efficiently, the programs for alignment problems over text segments consisting of a few sentences are relatively small and can be easily tackled with highly optimized general-purpose solvers.[5]

First, we define indicator variables for all potential phrase and arc edits in an alignment, as well as indicators that denote which pairs of tokens are aligned.

- $y^s_{ij\sim kl} \in \{0,1\}$ represents an alignment between the token sequence $t^s_{i\ldots j}$ from $T_s$ and $t^{s'}_{k\ldots l}$ from $T_{s'}$. We use $s'$ as shorthand for the segment index other than $s$, i.e., $s' = 3 - s$. Note that $y^s_{ij\sim kl}$ and $y^{s'}_{kl\sim ij}$ are equivalent for a given $i, j, k, l$ and refer to the same indicator.
- $\bar{y}^s_{ij} \in \{0,1\}$ represents a non-alignment or deletion of the token sequence $t^s_{i\ldots j}$ from either segment $T_s$.
- $z^s_{ij\sim kl} \in \{0,1\}$ represents an alignment between the dependency $d^s_{ij} \in D_s$ and $d^{s'}_{kl} \in D_{s'}$. Note that $z^s_{ij\sim kl}$ and $z^{s'}_{kl\sim ij}$ are equivalent for a given $i, j, k, l$ and refer to the same indicator.
- $\bar{z}^s_{ij} \in \{0,1\}$ represents a non-alignment or deletion of the dependency $d^s_{ij} \in D_s$.
- Finally, $x^s_{p\sim q} \in \{0,1\}$ indicates whether the token $t^s_p \in T_s$ participates in some phrase-based alignment with $t^{s'}_q \in T_{s'}$.

$$x^s_{p\sim q} = \begin{cases} 1, & \text{iff } \exists i, j, k, l \text{ s.t. } y^s_{ij\sim kl} = 1,\ i \le p \le j,\ k \le q \le l \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

[5] We use Gurobi: `http://www.gurobi.com`

Now, finding the optimal alignment between any sentence pair $\langle T_1, T_2\rangle$ is equivalent to solving the following optimization problem over the edit indicator variables:

$$\begin{aligned}\max_{y,z} \; & \sum_{i=1}^{n_1} \sum_{j=i}^{\min(n_1,i+\lambda)} \sum_{k=1}^{n_2} \sum_{l=k}^{\min(n_2,k+\lambda)} y_{ij\sim kl}\; \alpha_{\text{phr}}(\langle t^1_{i\ldots j}, t^2_{k\ldots l}\rangle) \\ & + \sum_{\substack{i,j:\\ d^1_{ij}\in D_1}} \sum_{\substack{k,l:\\ d^2_{kl}\in D_2}} z_{ij\sim kl}\; \alpha_{\text{arc}}(\langle d^1_{ij}, d^2_{kl}\rangle) \\ & + \sum_{s\in\{1,2\}} \left( \sum_{i=1}^{n_s} \sum_{j=i}^{\min(n_s,i+\lambda)} \bar{y}^s_{ij}\; \delta_{\text{phr}}(t^s_{i\ldots j}) \;+\; \sum_{d^s_{ij}\in D_s} \bar{z}^s_{ij}\; \delta_{\text{arc}}(d^s_{ij}) \right) \end{aligned} \tag{3}$$

where the parameter $\lambda$ controls the maximum number of tokens permitted in a phrase for alignment. The optimization problem requires some linear constraints in order to specify a complete and consistent alignment. The following constraints are applied for all $i = 1\ldots n_s$, $j = i\ldots \min(n_s, i+\lambda)$, $k = 1\ldots n_{s'}$, and $l = k\ldots \min(n_{s'}, k+\lambda)$ where $s \in \{1,2\}$.

1. Exactly one phrase edit must be active per token, ensuring a consistent segmentation for the phrase-based solution. Similarly, only one arc edit can be active per dependency.

$$\sum_{\substack{i,j:\\ i\le p\le j}} \sum_{k,l} y^s_{ij\sim kl} + \bar{y}^s_{ij} = 1 \qquad \forall p \in 1\ldots n_s \tag{4}$$

$$\sum_{k,l} z^s_{ij\sim kl} + \bar{z}^s_{ij} = 1 \qquad \forall i,j,k,l \text{ s.t. } d^s_{ij}\in D_s,\; d^{s'}_{kl}\in D_{s'} \tag{5}$$

2. An activated token pair indicator must participate in exactly one phrase alignment.

$$\sum_{\substack{i,j:\\ i\le p\le j}} \sum_{\substack{k,l:\\ k\le q\le l}} y^s_{ij\sim kl} = x^s_{p\sim q} \qquad \forall p\in 1\ldots n_s,\; q\in 1\ldots n_{s'} \tag{6}$$

3. In order to ensure that the phrase-based solution is consistent with the arc-based solution, arc alignments must activate corresponding token-pair alignment indicators.

$$z^s_{ij\sim kl} \le x^s_{i\sim k} \qquad \forall i,j,k,l \in 1,\ldots n_s \tag{7}$$

$$z^s_{ij\sim kl} \le x^s_{j\sim l} \qquad \forall i,j,k,l \in 1,\ldots n_s \tag{8}$$

4. If the governor and dependent of a dependency arc in one sentence are aligned to those of an arc in the other sentence, the corresponding arc alignment must be active.

$$x^s_{i\sim k} + x^s_{j\sim l} \le z^s_{ij\sim kl} + 1 \qquad \forall i,j,k,l \text{ s.t. } d^s_{ij}\in D_s,\; d^{s'}_{kl}\in D_{s'} \tag{9}$$

## 6 Experiments

We trained models with and without the dependency features using 20 epochs of averaged perceptron learning. Separate models were trained on the training corpus with just the SURE alignments and with the SURE+POSSIBLE alignments.[6] We used the unconstrained approach of Thadani and McKeown (2011) as a phrase-based baseline; this is an extension of MacCartney

[6] Note that all alignments are considered equally when evaluating on the SURE+POSSIBLE alignments.

| Alignments | System | Prec% | Rec% | $F_1$% | Exact% |
|---|---|---|---|---|---|
| Tokens/SURE | Meteor | 81.82 | 71.90 | 75.49 | 11.22 |
| | Phrase-based | 74.83 | 83.25 | **77.85** | 12.21 |
| | Phrase+Arc | 76.57 | 83.79 | **79.20** | 12.21 |
| Tokens/SURE+POSSIBLE | Meteor | 85.40 | 64.76 | 72.32 | 10.56 |
| | Phrase-based | 70.84 | 82.54 | **75.37** | 13.53 |
| | Phrase+Arc | 73.03 | 84.60 | **77.57** | 14.85 |
| Deps/SURE | Meteor | 84.64 | 58.03 | 65.60 | 17.49 |
| | Phrase-based | 76.07 | 78.42 | **75.10** | 23.10 |
| | Phrase+Arc | 73.56 | 84.27 | **76.30** | 20.79 |
| Deps/SURE+POSSIBLE | Meteor | 91.19 | 51.80 | 62.57 | 12.87 |
| | Phrase-based | 80.09 | 80.74 | **78.79** | 22.11 |
| | Phrase+Arc | 77.04 | 88.76 | **80.92** | 22.44 |

Table 1: Test set macro-averaged results on token alignments and projected dependency alignments over Stanford parses. $F_1$ increases are statistically significant in each case (see text).

et al. (2008) which outperforms a number of other alignment techniques (Och and Ney, 2003; Liang et al., 2006; Chambers et al., 2007). As an additional baseline, we ran Meteor on the test corpus using its precision-focused *max accuracy* setting, which we found to yield higher F-measure on the training corpus than the *max coverage* option. Table 1 shows the results.

It is evident that the feature-based aligners have much higher recall than Meteor, with some unsurprising loss in precision due to the conservative *max accuracy* matching. Compellingly, the joint model increases both precision and recall on aligned tokens over the phrasal model, with greater increases using the SURE+POSSIBLE alignments as expected. Jointly aligning arcs also helps considerably in recovering the dependency alignments projected onto Stanford parses from the gold standard phrase alignments. Wilcoxon signed-rank tests on $F_1$ indicate that all increases are statistically significant, with $p < 0.001$ in all cases except one, namely the increase on the SURE syntactic dependencies of the joint model over the phrasal model, where $p < 0.05$.

## Conclusion

We have presented a monolingual alignment strategy that jointly produces phrasal and syntactic dependency alignments using a discriminative structured prediction framework and an exact inference technique using ILP. Our alignment technique shows significant gains over recent phrase-based aligners and alignments obtained via the well-known Meteor metric. In future work, we intend to apply joint alignment approaches to additional corpora and develop more powerful similarity features over phrases and arcs.


## Acknowledgments

This work was supported in part by the Air Force Research Laboratory under a subcontract to FA8750-09-C-0179 and in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

**Disclaimer:** The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

## References


Barzilay, R. and Lee, L. (2003). Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*, NAACL '03, pages 16–23.

Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Bouamor, H., Max, A., and Vilnat, A. (2011). Monolingual alignment by edit rate computation on sentential paraphrase pairs. In *Proceedings of ACL-HLT*, pages 395–400.

Chambers, N., Cer, D., Grenager, T., Hall, D., Kiddon, C., MacCartney, B., de Marneffe, M.-C., Ramage, D., Yeh, E., and Manning, C. D. (2007). Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170.

Chang, M.-W., Goldwasser, D., Roth, D., and Srikumar, V. (2010). Discriminative learning over constrained latent representations. In *Proceedings of HLT-NAACL*, HLT '10, pages 429–437.

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

Collins, M. (2002). Discriminative training methods for hidden Markov models. In *Proceedings of EMNLP*, pages 1–8.

Dagan, I., Glickman, O., and Magnini, B. (2005). The pascal recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

DeNero, J. and Klein, D. (2008). The complexity of phrase alignment problems. In *Proceedings of ACL-HLT*, pages 25–28.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

Heilman, M. and Madnani, N. (2012). ETS: Discriminative edit models for paraphrase scoring. In *Proceedings of *SEM: The First Joint Conference on Lexical and Computational Semantics*, pages 529–535.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of HLT-NAACL*, HLT-NAACL '06, pages 104–111.

Lita, L. V., Ittycheriah, A., Roukos, S., and Kambhatla, N. (2003). tRuEcasIng. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

MacCartney, B., Galley, M., and Manning, C. D. (2008). A phrase-based alignment model for natural language inference. In *Proceedings of EMNLP*, EMNLP '08, pages 802–811.

MacCartney, B., Grenager, T., de Marneffe, M.-C., Cer, D., and Manning, C. D. (2006). Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 41–48.

Madnani, N., Tetreault, J., and Chodorow, M. (2012). Re-examining machine translation metrics for paraphrase identification. In *Proceedings of HLT-NAACL*, pages 182–190.

Marsi, E. and Krahmer, E. (2005). Explorations in sentence fusion. In *Proceedings of the European Workshop on Natural Language Generation*, pages 109–117.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

Quirk, C., Brockett, C., and Dolan, W. B. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, pages 142–149.

Thadani, K. and McKeown, K. (2008). A framework for identifying textual redundancy. In *Proceedings of COLING*, pages 873–880.

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

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of COLING*, COLING '96, pages 836–841.