Semi-Markov Phrase-based Monolingual Alignment

Xuchen Yao and **Benjamin Van Durme**
Johns Hopkins University
Baltimore, MD, USA

Chris Callison-Burch*
University of Pennsylvania
Philadelphia, PA, USA

Peter Clark
Allen Institute for Artificial Intelligence
Seattle, WA, USA

Abstract

We introduce a novel discriminative model for phrase-based monolingual alignment using a semi-Markov CRF. Our model achieves state-of-the-art alignment accuracy on two phrase-based alignment datasets (RTE and paraphrase), while doing significantly better than other strong baselines in both non-identical alignment and phrase-only alignment. Additional experiments highlight the potential benefit of our alignment model to RTE, paraphrase identification and question answering, where even a naive application of our model's alignment score approaches the state of the art.

1 Introduction

Various NLP tasks can be treated as an alignment problem: machine translation (aligning words in one language with words in another language), question answering (aligning question words with the answer phrase), textual entailment recognition (aligning premise with hypothesis), paraphrase detection (aligning semantically equivalent words), etc. Even though most of these tasks involve only a single language, alignment research has primarily focused on the bilingual setting (i.e., machine translation) rather than monolingual. Moreover, most work has considered token-based approaches over phrase-based.[1] Here we seek to address this imbalance by proposing better phrase-based models for monolingual word alignment.

Most token-based alignment models can extrinsically handle phrase-based alignment to some extent. For instance, in the case of `NYC` aligning to `New York City`, the single source word `NYC` may align three times separately to the target words: `NYC`↔`New`, `NYC`↔`York`, `NYC`↔`City`. Or in the case of identical alignment, `New York City` aligning to `New York City` is simply `New`↔`New`, `York`↔`York`, `City`↔`City`. However, it is not as clear how to token-align `New York` (as a city) with `New York City`. The problem is more prominent when aligning phrasal paraphrases or multiword expressions, such as `pass away` and `kick the bucket`. This suggests an intrinsically phrase-based alignment model.

The token aligner jacana-align (Yao et al., 2013a) has achieved state-of-the-art result on the task of monolingual alignment, based on previous work of Blunsom and Cohn (2006). It employs a Conditional Random Field (Lafferty et al., 2001) to align tokens from the source sentence to tokens in the target sentence, by treating source tokens as "observation" and target tokens as "hidden states". However, it is not designed to handle phrase-based alignment, largely due to the Markov nature of the underlying model: a state can only span one token each time, making it unable to align multiple consecutive tokens (i.e. a phrase). We extend this model by introducing semi-Markov states for phrase-based alignment: a state can instead span multiple consecutive time steps, thus aligning phrases on the source side. Also, we merge phrases on the target side to phrasal states, allowing the model to align phrases on the target side as well. We evaluate the resulting semi-Markov

*Performed while faculty at Johns Hopkins University.

[1] In this paper we use the term token-based alignment for one-to-one alignment and phrase-based for non one-to-one alignment, and word alignment in general for both.

CRF model on the task of phrase-based alignment, and then show a basic application in the NLP tasks of recognizing textual entailment, paraphrase identification, and question answering sentence ranking. The final phrase-based aligner is open-source.[2]

2 Related Work

Most work in monolingual alignment employs dependency tree/graph matching algorithms, including tree edit distance (Punyakanok et al., 2004; Kouylekov and Magnini, 2005; Heilman and Smith, 2010; Yao et al., 2013b), Particle Swarm Optimization (Mehdad, 2009), linear regression/classification models (Chambers et al., 2007; Wang and Manning, 2010), and min-cut (Roth and Frank, 2012). These works inherently only support token-based alignment, with phrase-like alignment achieved by first merging tokens to phrases as a *preprocessing* step.

The MANLI aligner (MacCartney et al., 2008) and its derivations (Thadani and McKeown, 2011; Thadani et al., 2012) are the first known phrase-based aligners specifically designed for aligning English sentence pairs. It applies discriminative perceptron learning with various features and handles phrase-based alignment of arbitrary phrase lengths. MANLI suffers from slow decoding time due to its large search space. This was optimized by Thadani and McKeown (2011) through Integer Linear Programming (ILP), where benefiting from modern ILP solvers they showed an order-of-magnitude speedup in decoding. Also, various syntactic constraints can be easily added, significantly improving exact alignment match rate for whole sentence pairs. Besides the common application of textual entailment and question answering, monolingual alignment has also been applied in the field of text generation (Barzilay and Lee, 2003; Pang et al., 2003).

Word alignment has been more explored in machine translation. The IBM models (Brown et al., 1993) allow many-to-one alignment and are essentially asymmetric. Phrase-based MT historically relied on heuristics (Koehn, 2010) to merge two sets of word alignment in opposite directions to yield phrasal alignment. Later, researchers explored non-heuristic phrase-based methods. Among them, Marcu and Wong (2002) described a joint probability model that generates both the source and target sentences simultaneously. All possible pairs of phrases in both sentences are enumerated and then pruned with statistical evidence. Deng and Byrne (2008) explored token-to-phrase alignment based on HMM models (Vogel et al., 1996) by explicitly modeling the token-to-phrase probability and phrase lengths. However, the token-to-phrase alignment is only in one direction: each target state still only spans one source word, and thus alignment on the source side is limited to tokens. Andrés-Ferrer and Juan (2009) extended the HMM-based method to Hidden Semi-Markov Models (HSMM) (Ostendorf et al., 1996), allowing phrasal alignments on the source side. Finally, Bansal et al. (2011) unified the HSMM models with the alignment by agreement framework (Liang et al., 2006), achieving phrasal alignment that agreed in both directions.

Despite successful usage of generative semi-Markov models in bilingual alignment, this has not been followed with models in discriminative monolingual alignment. Essentially monolingual alignment would benefit more from discriminative models with various feature extractions (just like those defined in MANLI) than generative models without any predefined feature (just like how they were used in bilingual alignment). To combine the strengths of both semi-Markov models and discriminative training, we propose to use the semi-Markov Conditional Random Field (Sarawagi and Cohen, 2004), which was first used in information extraction to tag continuous segments of input sequences and outperformed conventional CRFs in the task of named entity recognition. We describe this model in the following section.

3 The Alignment Model

Our objective is to define a model that supports phrase-based alignment of arbitrary phrase length. In this section we first describe a regular CRF model that supports one-to-one token-based alignment (Blunsom and Cohn, 2006; Yao et al., 2013a), then extend it to phrase-based alignment with the semi-Markov model.

[2] http://code.google.com/p/jacana/

3.1 Token-based Model

Given a source sentence $\mathbf{s}$ of length M, and a target sentence $\mathbf{t}$ of length N, the alignment from $\mathbf{s}$ to $\mathbf{t}$ is a sequence of target word indices $\mathbf{a}$, where $a_{i\in[1,M]} \in [0, N]$. We specify that when $a_i = 0$, source word s_i is aligned to a NULL state, i.e., deleted. This models a many-to-one alignment from source to target: multiple source words can be aligned to the same target word, but not vice versa. One-to-many alignment can be obtained by running the aligner in the other direction. The probability of alignment sequence $\mathbf{a}$ conditioned on both $\mathbf{s}$ and $\mathbf{t}$ is then:

$$p(\mathbf{a} \mid \mathbf{s}, \mathbf{t}) = \frac{\exp(\sum_{i,k} \lambda_k f_k(a_{i-1}, a_i, \mathbf{s}, \mathbf{t}))}{Z(\mathbf{s}, \mathbf{t})}$$

This assumes a first-order Conditional Random Field (Lafferty et al., 2001). Since the word alignment task is evaluated over F_1, instead of directly optimizing it, we choose a much easier objective (Gimpel and Smith, 2010) and add a cost function to the normalizing function $Z(\mathbf{s}, \mathbf{t})$ in the denominator:

$$Z(\mathbf{s}, \mathbf{t}) = \sum_{\hat{\mathbf{a}}} \exp(\sum_{i,k} \lambda_k f_k(\hat{a}_{i-1}, \hat{a}_i, \mathbf{s}, \mathbf{t}) + cost(\mathbf{a_y}, \hat{\mathbf{a}}))$$

where $\mathbf{a_y}$ is the true alignments. $cost(\mathbf{a_y}, \hat{\mathbf{a}})$ can be viewed as special "features" that encourage decoding to be consistent with true labels. It is only computed during training in the denominator because in the numerator $cost(\mathbf{a_y}, \mathbf{a_y}) = 0$. Hamming cost is used in practice without learning the weights (i.e., uniform weights). The more inconsistence there is between $\mathbf{a_y}$ and $\hat{\mathbf{a}}$, the more penalized is the decoding sequence $\hat{\mathbf{a}}$ through the cost function.

3.2 Phrase-based Model

The token-based model supports $1 : 1$ alignment. We first extend it in the direction of $l_s : 1$, where a target state spans l_s words on the source side (l_s source words align to 1 target word). Then we extend it in the direction of $1 : l_t$, where l_t is the target phrase length a source word aligns to (1 source word aligns to l_t target words). The final combined

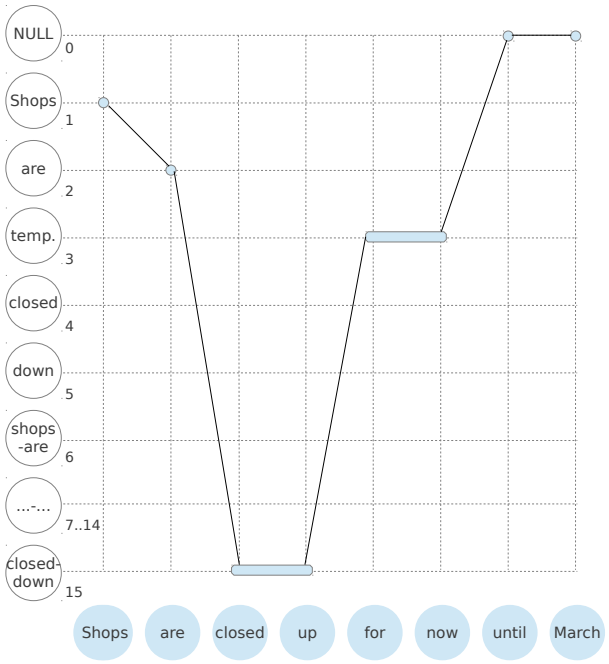

Figure 1: A semi-Markov phrase-based model example and the desired Viterbi decoding path. Shaded horizontal circles represent the source sentence (`Shops are closed up for now until March`) and hollow vertical circles represent the hidden states with state IDs for the target sentence (`Shops are temporarily closed down`). State 0, a NULL state, is designated for deletion. One state (e.g. state 3 and 15) can span multiple consecutive source words (a semi-Markov property) for aligning phrases on the source side. States with an ID larger than the target sentence length indicate "phrasal states" (states 6-15 in this example), where consecutive target tokens are merged for aligning phrases on the target side. Combining the semi-Markov property and phrasal states yields for instance, a 2×2 alignment between `closed up` in the source and `closed down` in the target.

model supports $l_s : l_t$ alignment. Throughout this section we use Figure 1 as an illustrative example, which shows phrasal alignment between the source sentence: (`Shops are closed up for now until March`) and the target sentence: (`Shops are temporarily closed down`).

$1 : 1$ alignment is a special case of $l_s : 1$ alignment where the target side state spans $l_s = 1$ source word, i.e., at *each* time step i, the source side word

s_i aligns to one state a_i and the next aligned state a_{i+1} only depends on the current state a_i. This is the Markovian property of the CRF. When $l_s > 1$, during the time frame $[i, i + l_s)$, all source words $[a_i, a_{i+l_s})$ share the same state a_i. Or in other words, the state a_i "spans" the following l_s time steps. The Markovian property still holds "outside" the time frame l_s, i.e., a_{i+l_s} still only depends on a_i, the previous state l_s time steps ago. But "within" the time frame l_s, the Markovian property does not hold any more: $[a_i, ..., a_{i+l_s-1}]$ are essentially the same state a_i. This is the semi-Markov property . States can be distinguished by this property into two types: *semi-Markovian states* and *Markovian states*.

We have generalized the regular CRF to a semi-Markov CRF. Now we define it by generalizing the feature function:

$$p(\mathbf{a} \mid \mathbf{s}, \mathbf{t}) = \frac{\exp(\sum_{i,k,l_s} \lambda_k f_k(a_{i-l_s}, a_i, \mathbf{s}, \mathbf{t}))}{Z(\mathbf{s}, \mathbf{t})}$$

At time i, the k-th feature function f_k mainly extracts features from the pair of source words $(s_{i-l_s}, ..., s_i]$ and target word t_{a_i} (still with a special case that $a_i = 0$ marks for deletion). Inference is still Viterbi-like: except for the fact during maximization, the Viterbi algorithm not only checks the previous *one* time step, but *all* l_s time steps. Suppose the allowed maximal source phrase length is L_s, define $V_i(a \mid \mathbf{s}, \mathbf{t})$ as the highest score along the decoding path until time i ending with state a:

$$V_i(a \mid \mathbf{s}, \mathbf{t}) = \max_{a_1, a_2, \ldots a_{i-1}} p(a_1, a_2, \ldots, a_i = a \mid \mathbf{s}, \mathbf{t})$$

then the recursive maximization is:

$$\begin{aligned} V_i(a \mid \mathbf{s}, \mathbf{t}) &= \max_{a'} \max_{l_s = 1 \ldots L_s} [V_{i-l_s}(a' \mid \mathbf{s}, \mathbf{t}) \\ &\quad + \Psi_i(a', a, l_s, \mathbf{s}, \mathbf{t})] \end{aligned}$$

with factor:

$$\Psi_i(a', a, l_s, \mathbf{s}, \mathbf{t}) = \sum_k \lambda_k f_k(a'_{i-l_s}, a_i, \mathbf{s}, \mathbf{t})$$

and the best alignment $\mathbf{a}$ can be obtained by backtracking the last state a_M from $V_M(a_M \mid \mathbf{s}, \mathbf{t})$.

Training a semi-Markov CRF is very similar to the inference, except for replacing maximization with summation. The forward-backward algorithm should also be used to dynamically compute the normalization function $Z(\mathbf{s}, \mathbf{t})$. Compared to regular CRFs, a semi-Markov CRF has a decoding time complexity of $O(L_s M N^2)$, a constant factor L_s (usually 3 or 4) slower.

To extend from $1:1$ alignment to $1:l_t$ alignment with one source word aligning to l_t target words, we simply explode the state space by L_t times with L_t the maximal allowed target phrase length. Thus the states can be represented as an $N \times L_t$ matrix. The state at (j, l_t) represents the target phrase $[t_j, ..., t_{j+l_t})$. In this paper we distinguish states by three types: NULL state ($j = 0, l_t = 0$), *token state* ($l_t = 1$) and *phrasal state* ($l_t > 1$).

To efficiently store and compute these states, we linearize the two dimensional matrix with a linear function mapping uniquely between the state ID and the target phrase offset/span. Suppose the target phrase t_j of length $l_{t_j} \in [1, L_t]$ holds a position $p_{t_j} \in [1, N]$, and the source word s_i is aligned to this state (p_{t_j}, l_{t_j}), a tuple for (position, span). Then state ID a_{s_i} is computed as:

$$a_{s_i}(p_{t_j}, l_{t_j}) = \begin{cases} p_{t_j} & l_{t_j} = 1 \\ N + (p_{t_j} - 1) \times L_t + l_{t_j} & 1 < l_{t_j} \leq L_t \end{cases}$$

Assume in Figure 1, $L_t = 2$, then the state ID for the phrasal state (5, 2) `closed-down` with $p_{t_j} = 5$ for the position of word `down` and $l_{t_j} = 2$ for the span of 2 words (looking "backward" from the word `down`) is: $5 + (5 - 1) \times 2 + 2 = 15$.

Similarly, given a state id a_{s_i}, the original target phrase position and length can be recovered through integer division and modulation. Thus during decoding, if one output state is 15, we would know that it uniquely comes from the phrasal state (5,2), representing the target phrase `closed down`.

This two dimensional definition of state space expands the number of states from $1 + N$ to $1 + L_t N$. Thus the decoding complexity becomes $O(M(L_t N)^2) = O(L_t^2 M N^2)$ with a usual value of 3 or 4 for L_t.

Now we have defined separately the $l_s : 1$ model and the $1 : l_t$ model. We can simply merge them to

have an $l_s : l_t$ alignment model. The semi-Markov property makes it possible for any target states to align phrases on the source side, while the two dimensional state mapping makes it possible for any source words to align phrases on the target side. For instance, in Figure 1, the phrasal state a_{15} represents the two-word phrase `closed down` on the target side, while still spanning for two words on the source side, allowing a 2×2 alignment. State a_{15} is phrasal, and at source word position 3 and 4 (spanning `closed up`) it is semi-Markovian. The final decoding complexity is $O(L_s L_t^2 M N^2)$, a factor of $30 \sim 60$ times slower than the token-based model (with a typical value of 3 or 4 for L_s and L_t).

In the following we describe features.

3.3 Feature Design

We reused features in the original token-based model based on string similarity, POS tags, position, WordNet, distortion and context. Then we used an additional chunker to mark phrase boundaries *only* for feature extraction:

Chunking Features are binary indicators of whether the phrase types of two phrases match. Also, we added indicators for mappings between source phrase types and target phrase types, such as "vp2np", meaning that a verb phrase in the source is mapped to a noun phrase in the target.

Moreover, we introduced the following lexical features:

PPDB Features (Ganitkevitch et al., 2013) include various similarity scores derived from a paraphrase database with 73 million phrasal and 8 million lexical paraphrases. Various paraphrase conditional probability was employed. For instance, for the ADJP/VP phrase pair `capable of` and `able to`, there are the following minus-log probabilities:

$$p(lhs|e1) = 0.1, p(lhs|e2) = 0.3, p(e1|lhs) = 5.0$$
$$p(e1|e2) = 1.3, p(e2|lhs) = 6.7, p(e2|e1) = 2.8$$
$$p(e1|e2, lhs) = 0.6, p(e2|e1, lhs) = 2.3$$

where $e1/e2$ are the phrase pair, and lhs is the left hand side syntactic non-terminal symbol. We did not use the syntactic part (e.g., `the` NP `of` NNS $\leftrightarrow$ `the` NNS `of` NP) of PPDB as we did not make the assumption that the input sentence pairs were well-formed (and newswire-like) English, or even of a language with a parser available. Also, for phrasal alignments, we ruled out those paraphrases spanning multiple syntactic structures, or of different syntactic structures (indicated as `[X]` in PPDB), for instance, `and crazy` $\leftrightarrow$ `, mad`.

Semantic Relatedness Feature is a single scaled number in $[0, 1]$ from the best performing system (Han et al., 2013) of the *Sem 2013 Semantic Textual Similarity (STS) task. We included this feature mainly to deal with cases where "related" words cannot be well measured by either paraphrases or distributional similarities. For instance, in one alignment dataset annotators aligned `married` with `wife`. Adding a few other words as comparison, the Han et al. (2013) system gives the following similarity scores:

```
married/wife:  0.85
married/husband:  0.84
married/child:  0.10
married/stone:  0.01
```

Name Phylogeny Feature (Andrews et al., 2012) is a similarity feature with a string transducer to model how one name evolves to another. Examples below show how similar is the name `Bill` associated with other names in log probability:

```
Bill/Bill:  -0.8
Bill/Billy:  -5.2
Bill/William:  -13.6
Bill/Mary:  -18.6
```

Finally, one decision we made during feature design was not to use any parsing-based features, with a permissive assumption that the input might not be well-formed English, or even not complete sentences (such as fragmented snippets from web search). The "deepest" linguistic processing stays at the level of tagging and chunking, making the model more easily extendable to other languages.

3.4 Feature Value

In this phrase-based model, the width of a state span over the source words depends on the competition between features fired on the phrases as a whole vs. the consecutive but individual tokens. We found it critical to assign feature values "fairly" among tokens and phrases to make sure that semi-Markov states and phrasal states fire up often enough for phrasal alignments.

	train	test	length	%align.
MSR06	800	800	29/11	36%
Edinburgh++	715	305	22/22	78%

Table 1: Statistics of the two manually aligned corpora, divided into training and test in sentence pairs. The length column shows average lengths of source and target sentences in a pair. %align. is the percentage of aligned tokens.

	1x1	1x2	1x3	2x2	2x3	3x3	more
MSR06	89.2	1.9	0.3	5.7	0.0	1.9	0.8
EDB++	81.9	3.5	0.8	8.3	0.4	3.0	2.1

Table 2: Percentage of various alignment sizes (undirectional, e.g., 1x2 and 2x1 are merged) after synthesizing phrasal alignment from token alignment in the *training* portion of two corpora.

To illustrate this in a simplified way, take `closed up↔closed down` in Figure 1, and assume the only feature is the normalized number of matching tokens in the pair. Then this feature firing on the following pairs would have values (the normalization factor is the maximal phrase length):

`closed↔closed`	1.0
`closed up↔closed`	0.5
`closed up↔up`	0.5
`closed up↔closed down`	0.5
`...↔...`	...

The desired alignment `closed up↔closed down` would not have survived the state competition due to its weak feature value. In this case the model would simply prefer a token alignment `closed↔closed` and up↔`...` (probably NULL).

Thus we upweighted feature values by the maximum source or target phrase length to encourage phrasal alignments, in this case `closed up ↔closed down`:1.0. Then this alignment would have a better chance to be picked out with additional features, such as with the PPDB and Semantic Relatedness Features, which are also upweighted by maximum phrase lengths.

4 Experiment

4.1 Data Preparation

There are two annotated datasets for training and testing. **MSR06**[3] (Brockett, 2007) has annotated alignments on the 2006 PASCAL RTE2 development and test corpora, with 1600 pairs in total. Semantically equivalent words and phrases in the premise and hypothesis sentences are aligned in a manner analogous to alignments in statistical machine translation. This dataset is asymmetric: on average the premises contain 29 words and the hypotheses 11 words. **Edinburgh++**[4] (Thadani et al., 2012) is a revised version of the Edinburgh paraphrase corpus(Cohn et al., 2008) with sentences from the following resources: 1. the Multiple-Translation Chinese corpus; 2. Jules Verne's novel Twenty Thousand Leagues Under the Sea. 3. the Microsoft Research paraphrase corpus (Dolan et al., 2004). The corpus is more balanced and symmetric: the source and target sentences are both 22 words long on average. Table 1 shows some statistics.

Both corpora contain mostly token-based alignment. For MSR06, MacCartney et al. (2008) showed that setting the allowable phrase size to be greater than one only increased F_1 by 0.2%. For Edinburgh++, the annotation guideline[5] explicitly instructs to "prefer smaller alignments whenever possible". Statistics shows that single token alignment counts 96% and 95% of total alignments in these two corpora separately. With such a heavy imbalance towards only token-based alignment, a phrase-based aligner would learn feature weights that award token alignments more than phrasal alignments.

Thus we synthesized phrasal alignments from continuous monotonic token alignments in these two corpora. We first ran the OpenNLP chunker through the corpora. Then for each phrase pair, if each token in the source phrase is aligned to a token in the target phrase in a monotonic way, and vice versa, we

[3] `http://www.cs.biu.ac.il/~nlp/files/RTE_2006_Aligned.zip`

[4] `http://www.ling.ohio-state.edu/~scott/#edinburgh-plusplus`

[5] `http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_guidelines.pdf`

merge these alignments to form one single phrasal alignment.[6] Table 2 lists the percentage of various alignment sizes after the merge. Two observations can be made: first, the portion of phrasal alignments increases to $10\% \sim 20\%$ after merging; second, allowing a maximal phrase length of 3 covers $98\% \sim 99\%$ of total alignments, thus a phrase length larger than 3 would be a bad trade-off for coverage vs speed.

4.2 Baselines and Evaluation Metrics

MacCartney et al. (2008) and Yao et al. (2013a) showed that the traditional MT bilingual aligner GIZA++ (Och and Ney, 2003) presented weak results on the task of monolingual alignment. Thus we instead used four other strong baselines:

Meteor (Denkowski and Lavie, 2011): a system for evaluating machine translation by aligning MT output with reference sentences. It is designed for the task of monolingual alignment and supports phrasal alignment. We used version 1.4 and default weights to optimize by maximum accuracy.

MANLI-constraint (Thadani and McKeown, 2011): a re-implemented MANLI system with ILP-powered decoding for speed and hard syntactic constraints to boost exact match rate, with reported numbers on MSR06.

MANLI-joint (Thadani et al., 2012): an improved version of MANLI-constraint that not only models phrasal alignments, but also alignments between dependency arcs, with reported numbers on the original Edinburgh paraphrase corpus.

jacana-token (Yao et al., 2013a): a token-based aligner with state-of-the-art performance on MSR06.

Note that the jacana-token aligner is open-source, so we were able to re-train it with exactly the same feature set used by our phrase-based model. This allows a fair comparison of model performance (token-based vs. phrase-based). The MANLI* systems are not available, thus we only reported their numbers from published papers.

The standard evaluation metrics for alignments are precision (P), recall (R), F_1, and exact matching rate (E) based on either tokens (two tokens are considered aligned iff they are aligned) or phrases (two tokens are considered aligned iff they are contained within phrases that are aligned). Following Thadani et al. (2012), we only report the results based on token alignments (which allows a partial credit if their containing phrases are not aligned), even for the phrase-based alignment task. The reasoning is that if a phrase-based aligner is already doing better than a token aligner in terms of token alignment scores, then the difference in terms of phrase alignment scores will be even larger. Thus showing the superiority of token alignment scores is sufficient.

4.3 Implementation and Training

The elements in the phrase-based model: dynamic state indices, semi-Markov and phrasal states, are not typically found in standard CRF implementations. Thus we implemented the phrase-based model in the Scala programming language, which is fully interoperable with Java, using one semi-Markov CRF package[7] as a reference. We used the L2 regularizer and LBFGS for optimization. OpenNLP[8] provided the POS tagger and chunker and JWNL[9] interfaced with WordNet (Fellbaum, 1998).

4.4 Results

Table 3 gives scores (in bigger fonts) of different aligners on MSR06 and Edinburgh++ and their corresponding phrasal versions. Overall, the token-based aligner did the best on the original corpora, in which single token alignment counts more than 95% of total alignment. The phrase-based aligner did slightly worse. We think the main reason was that it output more phrasal alignment, which in turn harms scores in token-based evaluation (for instance, if the gold alignment is `New`↔`New`, `York`↔`York`, then the phrasal alignment of `New York`↔`New York` would only have half the precision because it inherently also aligns `New` in the source with `York` in the target.). Further investigation showed that on the Edinburgh++ corpus, over-generated phrase-based alignment, when evaluated under just token alignment, contributed hurting about 1.1% of overall F_1,

[6] a few examples: `two Atlanta-based companies`↔`two Atlanta companies`, `the UK`↔`the UK`, `the 17-year-old`↔`the teenager`, `was held`↔`was held`.

[7] `http://crf.sf.net`

[8] `http://opennlp.apache.org/`

[9] `http://jwordnet.sf.net/`

a gap that would make the phrase aligner (85.9%) outperform the token aligner (86.4%).

On the phrasal alignment corpora (represented by MSR06P and EDB++P in Table 3), the phrase-based aligner did significantly better. Note that the overall F_1 and exact match rate are still much lower than those scores obtained from the original corpora, suggesting that the phrasal corpora present a much harder task. Furthermore, as a more "fair" comparison between the two aligners, we synthesized phrasal alignments from the output of the token-based aligner, just as how the phrased-based corpora were prepared, then evaluated its performance again. Still, on the EDB++P corpus, the token aligner was about 1.6% (current difference is 69.1% vs. 72.8%) worse than the phrase-based aligner.

Also, we want to emphasize that since the token-based aligner and the phrase-based aligner shared exactly the same features and lexical resources, the performance boost of the phrase-based aligner on the phrasal corpora results from a better model design: it is the semi-Markov property and phrasal states making the phrase-based aligner better.

To further investigate the performance of aligners with respect to different types of alignment, we divided the scores into those for identical alignments (such as `New`↔`New`) and non-identical alignments (such as `wife`↔`spouse`), indicated by the subscripts *i* and *n* in Table 3. In terms of identical alignment, most aligners were able to score more than 90%, but for non-identical alignment there was noticeable decrease. Still, on the phrasal alignment corpora, the phrase-based model has a much larger recall score for non-identical alignment than others.

We also divided scores with respect to token-only alignment and phrase-only alignment. Due to space limit, we only show results on synthesized Edinburgh++, in Table 4. Meteor and the token aligner inherently have either very limited or no support for phrasal alignment, thus they had very low scores on phrase-only alignment. We then ran the aligners in two directions and merged the results with the "union" MT heuristic to get better phrase support. But that still did not bring F_{1p}'s up to over 5%.

The phrase-based aligner baseline Meteor did worse than our aligners. We think there are two reasons: First, Meteor was not trained on these corpora. Second, Meteor only does strict word, stem, synonym and paraphrase matching but does not use any string similarity measures; this can be supported by the large difference between, for instance, F_{1i} and F_{1n}. In general Meteor did well on identical alignment, but not so well on non-identical alignment.

	System	P% P_i/P_n	R% R_i/R_n	F1% F_{1i}/F_{1n}	E%
MSR06 (78.6%)	Meteor	82.5 89.9/39.9	81.2 97.3/24.6	81.9 93.5/30.5	15.0
	MANLI-cons.	89.5	**86.2**	87.8	**33.0**
	token	**93.6** 96.6/77.7	83.5 96.9/35.6	**88.3** 96.8/48.8	32.1
	phrase	92.1 95.7/65.0	82.8 95.9/34.7	86.8 95.8/45.2	29.1
MSR06P (59.0%)	Meteor	82.5 89.9/40.1	68.3 97.3/8.8	74.7 93.5/14.5	7.3
	token	**92.9** 95.5/77.5	66.1 94.3/11.1	77.2 94.9/19.5	13.5
	phrase	83.5 94.9/55.5	**77.0** 94.2/48.1	**80.1** 94.5/51.5	**14.3**
EDB++ (75.2%)	Meteor	88.3 94.0/61.4	80.5 97.8/24.1	84.2 95.9/34.7	12.7
	MANLI-jnt*	76.6	**83.8**	79.2	12.2
	token	**91.3** 96.4/63.9	82.0 97.4/36.4	**86.4** 96.9/46.4	**15.0**
	phrase	90.4 96.0/57.4	81.9 97.8/38.3	85.9 96.9/46.0	13.7
EDB++P (51.7%)	Meteor	88.4 94.0/61.9	60.6 97.0/6.5	71.9 95.5/11.7	**2.9**
	token	**90.7** 96.2/58.6	55.8 91.3/7.1	69.1 93.7/12.7	2.3
	phrase	82.3 95.6/60.4	**65.3** 93.1/34.3	**72.8** 94.4/43.8	1.6

Table 3: Results on original (mostly token) and phrasal (P) alignment corpora, where (x%) indicates how much alignment is identical alignment, such as `New`↔`New`. E% stands for exact (perfect) match rate. Subscript *i* stands for corresponding scores for "identical" alignment and *n* for "non-identical". *: scores of MANLI-joint were for the original Edinburgh corpus instead of Edinburgh++ (with hand corrections) so it is not a direct comparison.

5 Applications

Natural language alignment can be applied to various NLP tasks. While how to most effectively apply

	System	P% P_t/P_p	R% R_t/R_p	F1% F_{1t}/F_{1p}	E%
EDB++P	Meteor	88.4 59.5/14.9	60.6 90.6/1.1	71.9 71.8/2.0	**2.9**
	token	**90.7** 59.4/21.4	55.8 85.5/0.9	69.1 70.1/1.7	2.3
	phrase	82.3 73.3/48.0	**65.3** 73.5/44.2	**72.8** 73.4/46.0	1.6

Table 4: Same results on the phrasal Edinburgh++ corpus but with scores divided by token-only alignment (subscript t) and phrase-only alignment (subscript p).

it is another topic, we simply show in this section using *just* alignment scores in binary prediction problems. Specifically, we pick the tasks of recognizing textual entailment (RTE), paraphrase identification (PP), and question answering sentence ranking (QA) described in Heilman and Smith (2010):

RTE: predicting whether a hypothesis can be inferred from the premise, with training data from RTE-1/2 and RTE-3 dev, and test from RTE-3 test.

PP: predicting whether two sentences are paraphrases, with training and test data from the MSR Paraphrase Corpus (Dolan et al., 2004).

QA: predicting whether a sentence contains the answer to the question, with training data from TREC-8 to TREC-12 and test data from TREC-13.

For each aligned pair, we can compute a normalized decoding score. Following MacCartney et al. (2008), we select a threshold score and predict true if the decoding score is above this threshold. For the tasks of RTE and PP, we tuned this threshold w.r.t the maximal accuracy on the training set, then reported performance on the test set. For the task of QA, since the evaluation methods in Mean Average Precision and Mean Reciprocal Rank only need a ranked list of answer sentences, and the scores on the test set are sufficient to provide the ranking, we did not tune anything on training but instead directly ran the aligner on the test set. All three tasks shared the same aligner model trained on the superset of MSR06 and Edinburgh++. Results are reported in Table 5. We could not report on Meteor as Meteor does not explicitly output alignment scores.

We did not expect the aligners to beat any of the

system	A%	P%	R%
de Marneffe et al. (2006)	60.5	61.8	60.2
MacCartney and Manning (2008)	**64.3**	**65.5**	63.9
Heilman and Smith (2010)	62.8	61.9	**71.2**
the token aligner	59.1	61.2	55.4
our phrasal aligner	57.6	57.2	68.8

(a) Recognizing Textual Entailment

system	A%	P%	R%
Wan et al. (2006)	**75.6**	**77**	90
Das and Smith (2009)	73.9	74.9	91.3
Heilman and Smith (2010)	73.2	75.7	87.8
the token aligner	70.0	72.6	88.1
our phrasal aligner	68.1	68.6	**95.8**

(b) Paraphrase Identification

system	MAP	MRR
Cui et al. (2005)	0.4271	0.5259
Wang et al. (2007)	0.6029	0.6852
Heilman and Smith (2010)	0.6091	0.6917
Yao et al. (2013b)	**0.6307**	**0.7477**
the token aligner	0.5982	0.6582
our phrasal aligner	0.6165	0.7333

(c) Question Answering Sentence Ranking

Table 5: Results (Accuracy, Precision, Recall, Mean Average Precision, Mean Reciprocal Rank) on the tasks of RTE, PP and QA.

state-of-the-art result since no sophisticated models were additionally used but only the alignment score. Still, the aligners showed competitive performance. It still follows the pattern from the alignment experiment that the phrasal aligner had higher recall and lower precision than the token aligner in the task of RTE and PP. In the QA task, the phrasal aligner performed better than all systems except for the top one.

6 Conclusion

We have introduced a phrase-to-phrase alignment model based on semi-Markov Conditional Random Fields. The combination of semi-Markov states and phrasal states makes phrasal alignment on both the source and target sides possible. The final phrase-

based aligner performed the best on two phrasal alignment corpora and showed its potential usage in three NLP tasks. Future work includes aligning discontinuous (gappy) phrases and integrating alignment more closely in NLP applications.

Acknowledgement

We thank Vulcan Inc. for funding this work. We also thank Jason Smith, Travis Wolfe and Frank Ferraro for various discussion, suggestion, comments and the three anonymous reviewers.

References

Jesús Andrés-Ferrer and Alfons Juan. 2009. A phrase-based hidden semi-markov approach to machine translation. In *Procedings of European Association for Machine Translation (EAMT)*, Barcelona, Spain, May. European Association for Machine Translation.

Nicholas Andrews, Jason Eisner, and Mark Dredze. 2012. Name phylogeny: a generative model of string variation. In *Proceedings of EMNLP 2012*.

Mohit Bansal, Chris Quirk, and Robert Moore. 2011. Gappy phrasal alignment by agreement. In *Proceedings of ACL*, Portland, Oregon, June.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL*, pages 16–23.

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of ACL2006*, pages 65–72.

Chris Brockett. 2007. Aligning the RTE 2006 corpus. Technical report, Microsoft Research.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

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 leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170.

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

Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan, and Tat-Seng Chua. 2005. Question answering passage retrieval using dependency relations. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 400–407, New York, NY, USA. ACM.

Dipanjan Das and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 468–476, Suntec, Singapore, August. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D Manning. 2006. Learning to distinguish valid textual entailments. In *Second Pascal RTE Challenge Workshop*.

Yonggang Deng and William Byrne. 2008. HMM word and phrase alignment for statistical machine translation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):494–507.

Michael Denkowski and Alon Lavie. 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*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of COLING*, Stroudsburg, PA, USA.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of NAACL-HLT*, pages 758–764.

Kevin Gimpel and Noah A. Smith. 2010. Softmax-margin CRFs: training log-linear models with cost functions. In *NAACL 2010*, pages 733–736.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC-EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*.

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

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.

Milen Kouylekov and Bernardo Magnini. 2005. Recognizing textual entailment with tree edit distance algorithms. In *PASCAL Challenges on RTE*, pages 17–20.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of NAACL*.

Bill MacCartney and Christopher D Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of ACL 2008*, pages 521–528.

Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of EMNLP*, pages 802–811.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP-2002*, pages 133–139.

Yashar Mehdad. 2009. Automatic cost estimation for tree edit distance using particle swarm optimization. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 289–292.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Mari Ostendorf, Vassilios V Digalakis, and Owen A Kimball. 1996. From HMM's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of NAACL*, pages 102–109.

Vasin Punyakanok, Dan Roth, and Wen T. Yih. 2004. Mapping Dependencies Trees: An Application to Question Answering. In *Proceedings of the 8th International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, Florida.

Michael Roth and Anette Frank. 2012. Aligning predicates across monolingual comparable texts using graph-based clustering. In *Proceedings of EMNLP-CoNLL*, pages 171–182, Jeju Island, Korea, July.

Sarawagi Sarawagi and William Cohen. 2004. Semimarkov conditional random fields for information extraction. *Advances in Neural Information Processing Systems*, 17:1185–1192.

Kapil Thadani and Kathleen McKeown. 2011. Optimal and syntactically-informed decoding for monolingual phrase-based alignment. In *Proceedings of ACL short*.

Kapil Thadani, Scott Martin, and Michael White. 2012. A joint phrasal and dependency model for paraphrase alignment. In *Proceedings of COLING 2012: Posters*, pages 1229–1238, Mumbai, India, December. The COLING 2012 Organizing Committee.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 836–841.

Stephen Wan, Mark Dras, Robert Dale, and Cécile Paris. 2006. Using dependency-based features to take the "para-farce" out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*.

Mengqiu Wang and Christopher D. Manning. 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of COLING*, pages 1164–1172, Stroudsburg, PA, USA.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. In *Proceedings of EMNLP-CoNLL*, pages 22–32, Prague, Czech Republic, June.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013a. A Lightweight and High Performance Monolingual Word Aligner. In *Proceedings of ACL 2013 short*, Sofia, Bulgaria.

Xuchen Yao, Benjamin Van Durme, Peter Clark, and Chris Callison-Burch. 2013b. Answer Extraction as Sequence Tagging with Tree Edit Distance. In *Proceedings of NAACL 2013*.