Multiple Alignment of Citation Sentences with Conditional Random Fields and Posterior Decoding

Ariel S. Schwartz[*]
EECS, Computer Science Division
UC Berkeley
Berkeley, CA 94720-1776
sariel@cs.berkeley.edu

Anna Divoli, Marti A. Hearst
School of Information
UC Berkeley
Berkeley, CA 94720-4600
{hearst,divoli}@ischool.berkeley.edu

Abstract

In scientific literature, sentences that cite related work can be a valuable resource for applications such as summarization, synonym identification, and entity extraction. In order to determine which equivalent entities are discussed in the various citation sentences, we propose *aligning* the words within these sentences according to semantic similarity. This problem is partly analogous to the problem of multiple sequence alignment in the biosciences, and is also closely related to the word alignment problem in statistical machine translation. In this paper we address the problem of multiple citation concept alignment by combining and modifying the CRF based pairwise word alignment system of Blunsom & Cohn (2006) and a posterior decoding based multiple sequence alignment algorithm of Schwartz & Pachter (2007). We evaluate the algorithm on hand-labeled data, achieving results that improve on a baseline.

1 Introduction

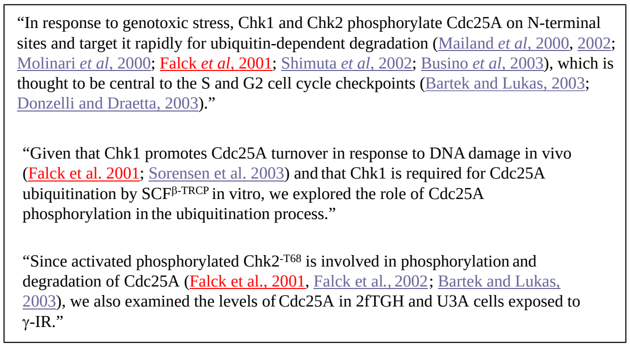
The scientific literature of biomedicine, genomics, and other biosciences is a rich, complex, and continually growing resource. With appropriate information extraction and retrieval tools, bioscience researchers can use the contents of the literature to further their research goals. With online full text of journal articles finally becoming the norm, new forms of citation analysis become possible.

Nearly every statement in biology articles is backed up by at least one citation, and, conversely, it is quite common for papers in the bioscience domain to be cited by 30–100 other papers. The cited facts are typically stated in a more concise way in the citing papers than in the original papers. Since the same facts are repeatedly stated in different ways in different papers, statistical models can be trained on existing citation sentences to identify similar facts in unseen text. Citation sentences also have the potential to be useful for text summarization and database curation. Figure 1 shows an example of three different citation sentences to the same target paper.

Most citation analysis work focuses on the citation network structure, to determine which papers are most central, or uses co-citation analysis to determine which papers are similar to one another in content (White, 2004; Liu, 1993; Garfield, 1955; Lipetz, 1965; Giles et al., 1998). In this paper we focus instead on analyzing the sentences that surround the citations to related work, which we termed *citances* in Nakov et al. (2004). In that paper we note that one subproblem for using citances for automated analysis is to identify the different concepts mentioned; a given paper may be cited for more than one fact or relation.

Citances often state similar information using varying words and phrases. In order to build concise summaries, those entities and relations that are expressed in different ways should be matched up, or *aligned*, so that subsequent processing steps will know what the core concepts are. In this paper we

[*] Current address: Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093-0412. Email: sariel@ucsd.edu.

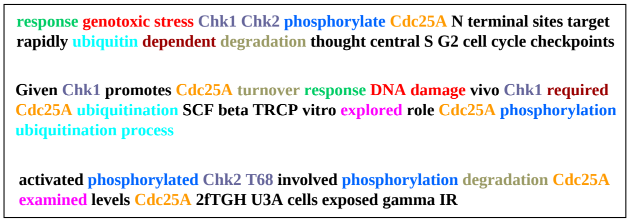


"In response to genotoxic stress, Chk1 and Chk2 phosphorylate Cdc25A on N-terminal sites and target it rapidly for ubiquitin-dependent degradation (Mailand *et al*, 2000, 2002; Molinari *et al*, 2000; Falck *et al*, 2001; Shimuta *et al*, 2002; Busino *et al*, 2003), which is thought to be central to the S and G2 cell cycle checkpoints (Bartek and Lukas, 2003; Donzelli and Draetta, 2003)."

"Given that Chk1 promotes Cdc25A turnover in response to DNA damage in vivo (Falck et al. 2001; Sorensen et al. 2003) and that Chk1 is required for Cdc25A ubiquitination by SCF$^{\beta\text{-TRCP}}$ in vitro, we explored the role of Cdc25A phosphorylation in the ubiquitination process."

"Since activated phosphorylated Chk2$^{\text{-T68}}$ is involved in phosphorylation and degradation of Cdc25A (Falck et al., 2001, Falck et al., 2002; Bartek and Lukas, 2003), we also examined the levels of Cdc25A in 2fTGH and U3A cells exposed to γ-IR."

Figure 1: **Example of three unaligned citances.**



response genotoxic stress Chk1 Chk2 phosphorylate Cdc25A N terminal sites target rapidly ubiquitin dependent degradation thought central S G2 cell cycle checkpoints

Given Chk1 promotes Cdc25A turnover response DNA damage vivo Chk1 required Cdc25A ubiquitination SCF beta TRCP vitro explored role Cdc25A phosphorylation ubiquitination process

activated phosphorylated Chk2 T68 involved phosphorylation degradation Cdc25A examined levels Cdc25A 2fTGH U3A cells exposed gamma IR

Figure 2: **Example of three normalized aligned citances.** Homologous entities are colored the same. Unaligned entities are black.

build on the work of Nakov et al. (2004) by tackling the entity normalization step.

The citance alignment problem is partially analogous to the problem of multiple alignment of biological sequences (Durbin et al., 1998). In both cases the goal is to *align* homologous entities that are derived from the same ancestral entity. While in biology homology is well-defined in the molecular level, in the citances case it is defined in the semantic level, which is much more subjective. Given a group of citances that cite the same target paper, we loosely define *semantic homology* as a symmetric, transitive, and reflexive relation between two entities (words or phrases) in the same or different citance that have similar semantics in the context of the cited paper.

Figure 1 shows an example of three citances that cite the same target paper (Falck et al., 2001). A multiple alignment of the entities in the same citances (after removal of stopwords) is shown in Figure 2. Homologous entities are colored the same. This small example illustrates some of the main challenges of *multiple citance alignment* (MCA). While orthographic similarity can help to identify semantic homology (e.g., *phosphorylate* and *phosphorylation*), it can also be misleading (e.g., *cell cycle* and *U3A cells*). In addition, semantic homology might not include any orthographic clues (e.g., *genotoxic stress* and *DNA damage*).

Unlike global multiple sequence alignment (MSA) in genomics, where each character can be aligned to at most one character in every other sequence, in multiple citance alignment, each word can be aligned to any number of words in other sentences. Another major difference between the two problems is the fact that while the sequential ordering of characters must be maintained in multiple sequence alignment, this is not the case for multiple citance alignment.

MCA is also related to the problem of word alignment in statistical machine translation (SMT) (Och and Ney, 2003). However, unlike SMT alignment, MCA aligns multiple citances in the same language rather than a pair of sentences in different languages.

In this paper we present an MCA algorithm that is based on an extension to the posterior decoding algorithm for MSA called AMAP (Schwartz et al., 2006; Schwartz and Pachter, 2007), with an underlying pairwise alignment model based on the CRF SMT alignment of Blunsom & Cohn (2006).

2 Multiple citance alignments

Let $\mathcal{G} \triangleq \{C^1, C^2, \ldots, C^K\}$ be a group of K citances that cite the same target paper, where the i^{th} citance is a sequence of words $C^i \triangleq C^i_1 C^i_2 \cdots C^i_{n^i}$, and $c^i \triangleq \{c^i_1, c^i_2, \ldots, c^i_{n^i}\}$ is the set of word indices of C^i. A *pairwise citance alignment* of C^i and C^j is an equivalence (symmetric, reflexive, and transitive) relation $\sim_{ij}$ on the set $c^i \cup c^j$. The expression $c^i_k \sim_{ij} c^j_l$ means that according to the pairwise alignment $\sim_{ij}$ word k in citance C^i and word l in citance C^j are aligned. A *multiple citance alignment* (MCA) is an equivalence relation $\sim \triangleq \left(\bigcup_{ij} \sim_{ij}\right)^+$ on the set $\bigcup_i c^i$, which is the transitive closure of the union of all pairwise alignments of citance pairs in $\mathcal{G}$. Taking the transitive closure and not only the union of all pairwise alignments ensures that the MCA is an equivalence relation.

An MCA $\sim$ defines a partition of the set of all word indices $c \triangleq \bigcup_{ik} \{c^i_k\}$, which is of size $n \triangleq$

$|c| = \sum_i n^i$. Therefore, the number of distinct MCAs of $\mathcal{G}$ is the number of partitions of a set of size n. This number is called the n^{th} Bell number (Rota, 1964)

$$B_n \triangleq \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}. \quad (1)$$

Asymptotically, B_n grows faster than an exponential but slower than a factorial. For example $B_{100} \approx 10^{116}$. Obviously, enumerating all possible MCAs is impractical even for small problems.

3 Probabilistic model for MCA

Unlike biological sequences, for which pair-HMMs are a natural choice for modeling evolutionary processes between two sequences, there is no simple generative model that can be used for modeling pairwise citance alignment. Most of the work on pairwise alignment of sentences at the word level has been done in the statistical machine translation (SMT) community.

Och & Ney (2003) present an overview and comparison of the most common models used for SMT word alignments. Out of the models they describe, the HMM models are the most expressive models that can compute posterior probabilities using the forward-backward algorithm. However, unlike sequence alignments, there are no ordering constraints in word alignments, and the alignments are many-to-many as opposed to one-to-one. Therefore, the SMT HMM models cannot be based on pair-HMMs, which generate two sentences simultaneously. Rather, they are directional models that model the probability of generating a target sentence given a source sentence. In other words they only model many-to-one alignments, recovering the many-to-many alignments in a preprocessing step. Therefore, SMT HMMs can only compute the posterior probabilities $P(c^i_k \rightsquigarrow c^j_l | C^i, C^j)$ and $P(c^j_l \rightsquigarrow c^i_k | C^i, C^j)$, where the relation $\rightsquigarrow$ represents the (directional) event that a source word is translated into a target word. Nevertheless, recently such posterior probabilities have been used in SMT word alignment system as an alternative to Viterbi decoding, and helped to improve the performance of such systems (Matusov et al., 2004; Liang et al., 2006).

Generative models like HMMs have several limitations. First, they require relatively large training data, which is difficult to attain in case of SMT word alignment, and even more so in the case of MCA. Second, generative models explicitly model the inter-dependence of different features, which reduces the ability to incorporate multiple arbitrary features into the model. Since orthographic similarity is not a strong enough indication for semantic homology in MCA, we would like to be able to incorporate multiple inter-dependent features into a single model, including orthographic, contextual, ontological, and lexical features.

Recently, several authors have described discriminative SMT alignment models (Moore, 2005; Lacoste-Julien et al., 2006; Blunsom and Cohn, 2006). However, to the best of our knowledge only the model of Blunsom & Cohn (2006), which is based on a Conditional Random Field (CRF) (Lafferty et al., 2001), can compute word indices pairs' directional posterior probabilities, like those computed by the HMM models. Therefore, we decided to adopt the CRF-based model to the MCA problem.

3.1 Conditional random fields for word alignment

The model of Blunsom & Cohn (2006) is based on a linear chain CRF, which can be viewed as the undirected version of an HMM. The CRF models a many-to-one pairwise alignment, in which every source word can get aligned to zero or one target words, but every word in the target sentence can be the target of multiple source words. CRFs define a conditional distribution over a latent labeling sequence given observation sequence(s). In the case of CRF for word alignment, the observed sequences are the source and target sentences (citances), and the latent labeling sequence is the mapping of source words to target word-indices. Given a source citance C^i of length n^i, and a target citance C^j of length n^j, the many-to-one alignment of C^i to C^j is the relation $\rightsquigarrow$. Since this is a many-to-one alignment, $\rightsquigarrow$ can be represented by a vector a of length n^i. The CRF models the probability of the alignment a conditioned on C^i and C^j as follows:

$$P_\Lambda(a|C^i, C^j) = \frac{\exp\left(\sum_t \sum_k \lambda_k f_k(t, a_{t-1}, a_t, C^i, C^j)\right)}{Z_\Lambda(C^i, C^j)}, \quad (2)$$

where $f \triangleq \{f_k\}$ are the model's features, $\Lambda \triangleq \{\lambda_k\}$ are the features' weights, and $Z_\Lambda(C^i, C^j)$ is the partition (normalization) function which is defined as:

$$Z_\Lambda(C^i, C^j) \triangleq \sum_a \exp\left(\sum_t \sum_k \lambda_k f_k(t, a_{t-1}, a_t, C^i, C^j)\right). \tag{3}$$

Parameters are estimated from fully observed data (manually aligned citances) using a maximum a posteriori estimate. The parameter estimation procedure is described in more details in the original paper. Blunsom & Cohn (2006) use Viterbi decoding to find an alignment of two sentences given a trained CRF model, $a^* \triangleq \operatorname{argmax}_a P_\Lambda(a|C^i, C^j)$. However, the posterior probabilities of the labels at each position can be calculated as well using the forward-backward algorithm:

$$P_\Lambda(c^i_l \rightsquigarrow c^j_k|C^i, C^j) = P_\Lambda(a_l = c^j_k|C^i, C^j) = \frac{\alpha_l(c^j_k|C^i, C^j)\beta_l(c^j_k|C^i, C^j)}{Z_\Lambda(C^i, C^j)} \tag{4}$$

where α_l and β_l are the forward and backward vectors that are computed with the forward-backward algorithm (Lafferty et al., 2001).

3.2 The posterior decoding algorithm for MCA

Ultimately, the success of an MCA algorithm should be judged by its effect on the success of the citance analysis systems that use MCAs as their input. However, measuring this effect directly is difficult, since high-level tasks such as summarization are difficult to evaluate objectively. More to the point, it is difficult to quantify the contribution of the MCA accuracy to the accuracy of the high-level system that uses it. A more practical alternative is to measure the accuracy of MCAs directly using a meaningful accuracy measure, under the simplifying assumption that there is a strong correlation between the measured MCA accuracy and the performance of the high-level application.

We argue that a useful utility function should be correlated (or even identical) to the accuracy measure used to evaluate the performance of an algorithm. In addition, the utility function should be easily decomposable, to enable direct optimization using posterior-decoding. Although any accuracy measure that is acceptable as a single performance measure can be used to guide the design of the utility function, metric-based accuracy measures have several noticeable advantages. First, a metric formalizes the intuitive notion of distance. Hence, an accuracy measure which is based on a metric follows the intuition that reducing the distance to the correct answer should increase the accuracy of the predicted answer. Therefore, defining a metric space for the objects of a given problem leads to a natural definition of accuracy. Another advantage of using a metric-based accuracy measure is the ability to provide bounds in the search space using the triangle inequality. For example, while searching for the answer with the optimal (metric-based) expected utility, a step of length x can only change the expected utility as well as the actual utility by at most $\pm x$ units. Examples of more complex bounds using metric loss functions are described in (Schlüter et al., 2005) and (Domingos, 2000).

Schwartz et al. (2006) define the *alignment metric accuracy* (AMA), which is a utility function for one-to-one MSA. Intuitively, AMA measures the fraction of characters that are aligned correctly according to the reference alignment, either to another character or to a gap (null). We extend the definition of AMA to the case of many-to-many MCA.

A good utility function for MCA should give partial credit to word positions that align to some of the correct word positions while penalizing for aligning to wrong word positions. To help define such a utility function we define the following. Let $m^{ij}_{\sim}(c^j_l) \triangleq \{c^i_k \in c^i | c^i_k \sim c^j_l\}$ be the set of all word positions in citance C^i that align to word position l in citance C^j according to MCA $\sim$. We can then define the following utility function for the MCA $\sim^p$ of the citance group $\mathcal{G}$ given a reference MCA $\sim^r$:

$$U_{AMA}(\sim^r, \sim^p) \triangleq \frac{\sum_{ijl|i \neq j} U_{set\_agreement}\left(m^{ij}_{\sim^r}(c^j_l), m^{ij}_{\sim^p}(c^j_l)\right)}{n(K-1)}, \tag{5}$$

where n is the number of word indices in $\mathcal{G}$, $K \triangleq |\mathcal{G}|$ is the number of citances in the group, and $U_{set\_agreement}$ is any utility function for agreement between sets that assigns values in the range $[0, 1]$.

$U_{set\_agreement}$ can be viewed as a "score" assigned to each word position based on the agreement between the two alignments with regards to the other word positions that align to it. Using a 0–1 loss as the set agreement score is equivalent to the original AMA. Other utility functions, such as Dice, Jaccard and Hamming can be used as $U_{set\_agreement}$. However, only metric-based utility functions will result in a metric-based U_{AMA} utility function. It is easy to see that $1 - U_{AMA}$ satisfies all the requirements of a metric, i.e., it is non-negative, equals zero if and only if $\sim^r = \sim^p$, symmetric, and obeys the triangle inequality, since if the triangle inequality holds for $1 - U_{set\_agreement}$, it must hold for a sum of $1 - U_{set\_agreement}$ values. (We refer the reader to Schwartz (2007) for a longer discussion of the properties of the different utility functions.) We define the AMA for MCA by setting the $U_{set\_agreement}$ to be the Braun-Blanquet coefficient (Braun-Blanquet, 1932), which is defined as:

$$U_{Braun-Blanquet}\left(m^{ij}_{\sim^r}(c^j_l), m^{ij}_{\sim^p}(c^j_l)\right) \triangleq \begin{cases} 1 & \text{if } m^{ij}_{\sim^r}(c^j_l) = \emptyset \\ & \text{and } m^{ij}_{\sim^p}(c^j_l) = \emptyset \\ \frac{|m^{ij}_{\sim^r}(c^j_l) \cap m^{ij}_{\sim^p}(c^j_l)|}{\max\left\{|m^{ij}_{\sim^r}(c^j_l)|, |m^{ij}_{\sim^p}(c^j_l)|\right\}} & \text{otherwise} \end{cases} . \tag{6}$$

Caillez & Kuntz (1996) show that the Braun-Blanquet coefficient is based on a metric.

As with the MSA case, a family of utility functions can be defined to enable control of the recall/precision trade-off. Unlike MSA, in the case of MCA two free parameters are needed, in order to have better control of the trade-off using posterior-decoding. In addition to a *gap-factor* that controls the threshold at which unaligned words start to get aligned, a *match-factor* is added to enable control of the number of word-positions each word aligns to. The result is the following utility function:

$$U_{\mu,\gamma}(\sim^r, \sim^p) \triangleq \frac{1}{n(K-1)} \sum_{ijl|i \neq j} \Bigg(\mu^{|m^{ij}_{\sim^p}(c^j_l)|} \frac{|m^{ij}_{\sim^r}(c^j_l) \cap m^{ij}_{\sim^p}(c^j_l)|}{\max\left\{|m^{ij}_{\sim^r}(c^j_l)|, |m^{ij}_{\sim^p}(c^j_l)|, 1\right\}} + \gamma \mathbf{1}\{m^{ij}_{\sim^r}(c^j_l) = m^{ij}_{\sim^p}(c^j_l) = \emptyset\} \Bigg), \tag{7}$$

where $\gamma \in [0, \infty)$ is a gap-factor, and $\mu \in (0, \infty)$ is a match factor. The neutral value for both parameters is 1. Increasing γ results in increased utility to sparser MCAs, while reducing γ increases the utility of denser alignments. However, in the case of MCA, the gap-factor only affects the first aligned word position, but it cannot affect the number of word positions each word is aligned to. The match-factor adds this functionality by rewarding MCAs that align words to multiple word positions when $\mu > 1$, and penalizing such MCAs when $\mu < 1$.

Given a group of K citances $\mathcal{G}$ and a trained CRF model, the goal of the MCA algorithm is to find the MCA $\sim^* \triangleq \operatorname{argmax}_{\sim^p} E_{\sim^t} U_{\mu,\gamma}(\sim^t, \sim^p)$ that maximizes the expected utility. Since searching the space of possible MCAs exhaustively is infeasible, we resort to a simple heuristic for predicting an MCA. Instead of searching for a global optimum, the predicted MCA is defined as the equivalence (symmetric transitive) closure of the union of multiple local optima. For each target word position c^j_l and every source citance C^i the combination of source word positions $c^i_\circ$ that maximize the expected set-agreement score of c^j_l is added to the predicted MCA. Formally, let $\mathcal{P}(c^i)$ be the power-set of c^i, then we define the predicted MCA as $\sim^p \triangleq \left(\leadsto^p \cup (\leadsto^p)^{-1}\right)^+$, where $\leadsto^p$ is defined as:

$$\leadsto^p \triangleq \bigcup_{ijl|i \neq j} \{c^j_l\} \times \operatorname*{argmax}_{c^i_\circ \in \mathcal{P}(c^i)} E_{m^{ij}_{\sim^t}(c^j_l)} \left(\mu^{|c^i_\circ|} \frac{|m^{ij}_{\sim^t}(c^j_l) \cap c^i_\circ|}{\max\left\{|m^{ij}_{\sim^t}(c^j_l)|, |c^i_\circ|, 1\right\}} + \gamma \mathbf{1}\{m^{ij}_{\sim^t}(c^j_l) = c^i_\circ = \emptyset\} \right). \tag{8}$$

The value of $\leadsto^p$ can be computed from the CRF directional posterior probabilities as follows:

$$\leadsto^p = \bigcup_{ijl|i \neq j} \{c^j_l\} \times \operatorname*{argmax}_{c^i_\circ \in \mathcal{P}(c^i)} \sum_{c^i_* \in \mathcal{P}(c^i)} P\left(m^{ij}_{\sim^t}(c^j_l) = c^i_*\right) \left(\mu^{|c^i_\circ|} \frac{|c^i_* \cap c^i_\circ|}{\max\{|c^i_*|, |c^i_\circ|, 1\}} + \gamma \mathbf{1}\{c^i_* = c^i_\circ = \emptyset\} \right), \tag{9}$$

and using an independence assumption we get:

$$\begin{aligned}&\leadsto^p \approx \bigcup_{ijl|i\neq j} \{c_l^j\} \times \operatorname*{argmax}_{c_\circ^i \in \mathcal{P}(c^i)} \sum_{c_*^i \in \mathcal{P}(c^i)} \\ &\left(\prod_{c_k^i} \Big(P_\Lambda(c_k^i \leadsto c_l^j | C^i, C^j) \mathbf{1}\{c_k^i \in c_*^i\} + \right. \\ &\left. (1 - P_\Lambda(c_k^i \leadsto c_l^j | C^i, C^j)) \mathbf{1}\{c_k^i \notin c_*^i\}\Big) \right) \\ &\left(\mu^{|c_\circ^i|} \frac{|c_*^i \cap c_\circ^i|}{\max\{|c_*^i|, |c_\circ^i|, 1\}} + \gamma \mathbf{1}\{c_*^i = c_\circ^i = \emptyset\}\right). \end{aligned} \tag{10}$$

Note that although the directional posterior probabilities are used to generate the predicted MCA, the result is a many-to-many alignment, since the union is done over all pairs of sequences in both directions. The calculation in Equation (10) can be computationally intensive in practice, as it requires $|\mathcal{P}(c^i)|^2 = 2^{2n^i}$ operations for each word position c_l^j and citance C^i. This can be overcome by restricting the combinations of source word positions (c_*^i and $c_\circ^i$) to include only the the top MAX_SOURCES source words with a minimum posterior probability of MIN_PROB to align to c_l^j ($P_\Lambda(c_k^i \leadsto c_l^j | C^i, C^j) \geq$ MIN_PROB). In our implementation we set MAX_SOURCES to 8 and MIN_PROB to 0.01. Additionally, the probabilities of each combination c_*^i can be calculated only once, since it is independent of $c_\circ^i$. This reduces the total computational complexity of calculating $\leadsto^p$ to $O\left(2^{16}(K^2 - K)\max_{n^i}\{n^i\}\right)$.

4 Data sets

Since citance alignment is a new task, we had to create our own evaluation and training sets. We restricted the domain of the target papers to molecular interactions, a domain which is actively researched in the biosciences text mining community (Hirschman et al., 2002). The biologist in our group annotated citances to 6 target papers. The training set consisted of 40 citances to 4 different target papers (10 citances each; we wanted to have variety in the training set). The development set consisted of 51 citances to the fifth target paper, and the test set contained 45 citances to the sixth target paper.

For each target paper we downloaded the full text of those papers citing it that were available in HTML format. The link structure of the cited references in the HTML documents allowed us to automatically extract citances to a given target paper. We defined a citance to be the full sentence that contains a citation to the target paper. Each citance was then tokenized, and normalized by removing all stopwords from a predefined list.

One goal of the annotation was to cover as much of the content of the citances as possible. Another goal was consistency; our biologist manually followed a small number of rules to determine semantic similarity. Distinct semantic units (words or phrases) were identified and given an annotation ID. Within each group of citances, words or phrases that share semantic similarity were annotated with the same ID.

Using the manually annotated citance groups, pairwise word alignments were generated for every source-target pair of citances from every group. That resulted in a training, development, and test sets of 180, 1275, and 990 pairwise alignments respectively.

Alignments that were used for development and testing were generated as many-to-many alignments. However, many-to-many alignments are not suitable for the training the many-to-one CRF alignment model. When a given source word c_k^i aligns to multiple words in the target citance, the CRF model chooses only one target word as a true positive, while incorrectly treating the other true positive target words as true negatives. To alleviate this problem, in such cases we replaced all true-positive target words other than the first with '*', thus forcing them to act as real true negatives for the purpose of training. This adjustment does not solve the inherent limitation of the CRF's many-to-one modeling of a many-to-many alignment, but it prevents learning incorrect weights for good features.

5 Feature engineering

The CRF alignment model can combine multiple overlapping features. We evaluated the effectiveness of different features by training models on the training set and evaluating their performance on the development set. We considered variations of features that were part of the original system of Blunsom & Cohn (2006), and also designed new features that are specific to the problem of MCA.

Orthographic features

We used the following orthographic features from the original system of Blunsom & Cohn (2006) (below all features are either Boolean indicator functions (b) or real valued (r)):

(b) exact string similarity of source-target words;
(b) every possible source-target pair of length 3 prefixes;
(b) exact string match of length 3 prefixes;
(b) exact string match of length 3 suffixes;
(r) absolute difference in word lengths;
(b) both words are shorter than 4 characters.

In addition, the following orthographic features were added: indicator that both words include capital letters, and normalized edit-similarity of the two words $(1 - \frac{edit\_distance(c_k^i, c_l^j)}{max\{|c_k^i|, |c_l^j|\}})$.

Markov features

We used the following Markov features from the original system:

(r) absolute jump width ($abs(a_t - a_{t-1} - 1)$, which measures the distance between the target words of adjacent source words;

(r) positive jump width ($max\{a_t - a_{t-1} - 1, 0\}$);

(r) negative jump width ($max\{a_{t-1} + 1 - a_t, 0\}$);

(b) transition from null aligned source-word to non-null aligned source-word;

(b) transition from non-null aligned source-word to null aligned source-word;

(b) transition from null aligned source-word to null aligned source-word.

In addition we added the following Markov features in order to model the tendency of certain words to be part of longer phrases:

(b) source-word aligns to the same target-word as the previous source-word;

(b) source-word aligns to the same target-word as the next source-word;

(b) transition from non-null aligned source-word to non-null aligned source-word.

Sentence position: We included the relative sentence position feature from the original system, which is defined as $abs(\frac{a_t}{|c^j|} - \frac{t}{c^i})$. Although it was not expected to be relevant for MCA, since the citances are not expected to align along the diagonal, this feature slightly improved the performance of the development set.

Null: An indicator function for leaving a source-word unaligned was retained from the original system. This is an essential feature since without it the CRF tends to over-align words, and produces meaningless posterior probabilities.

Ontological features: Orthographic and positional features alone do not cover all cases of semantic homology. We therefore included features that are based on domain specific ontologies.

Using an automated script we mapped specific words and phrases in every citance to MeSH[1] terms, Gene identifiers from Entrez Gene,[2] UniProt,[3] and OMIM.[4] We then added features indicating when the source and target words are annotated with the same MeSH term or the same gene identifier. We tried numerous features that compare MeSH terms based on their distance in the ontology, and other features that indicate whether a word is part of a longer term. However, none of these feature were selected for the final system.

In addition to biological ontologies we added a feature for semantic word similarity between the source and target words, based on the Lin (1998) WordNet similarity measure.

6 Results

We modified the CRF alignment system of Blunsom & Cohn (2006) to support MCA by incorporating the posterior decoding algorithm from Section 3.2 into the existing system. The CRF model was trained using the features that were selected using the development set, on a dataset that included the training and development MCAs. All the performance results in this section are reported on the test set, which includes 990 pairs of citances $(45 \times 44/2)$, with a total of 34188 words (8547×44). On average, 20% of the source words are aligned to at least one other target word in a given reference pairwise alignment. Since the union of all the pairwise alignments results in only a single test MCA, it is hard to make strong arguments about the performance

[1] http://www.nlm.nih.gov/mesh/

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[3] http://www.pir.uniprot.org/

[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

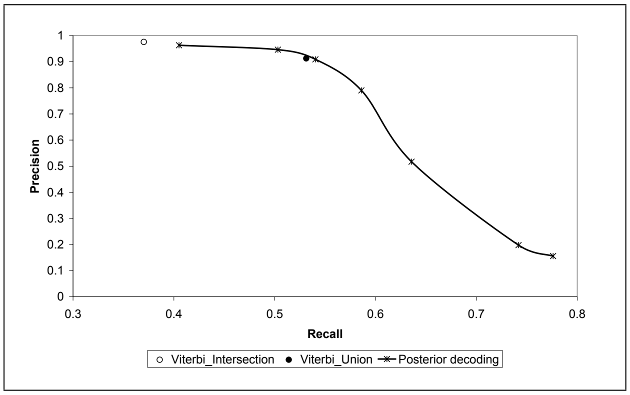


Figure 3: **Recall/Precision curve of pairwise citance alignments comparing Viterbi to posterior decoding.**

of the system in general. Therefore, we concentrate our discussion on general trends, and do not claim that the specific performance numbers we report here are statistically significant. As a point of comparison, the SMT community has been evaluating performance of word-alignment systems on an even smaller dataset of 447 pairs of non-overlapping sentences (Mihalcea and Pedersen, 2003).

We first analyze the performance of the system on pairwise citance alignments. Instead of taking the equivalence closure of $\leadsto^p$ we take only the symmetric closure. The result is 990 many-to-many pairwise alignments. In order to evaluate the effectiveness of the posterior-decoding algorithm, we generate the Viterbi alignments using the same CRF model. The Viterbi many-to-many pairwise alignments are then generated by combining equivalent pairs of many-to-one alignments using three different standard symmetrization methods for word-alignment—union, intersection, and the refined method of Och & Ney (2003).

Figure 3 shows the recall/precision trade-off of the pairwise posterior-decoding and Viterbi alignments. The curve for the posterior-decoding alignments was produced by varying the gap and match factors. For the Viterbi alignments, only three results could be generated (one for each symmetrization method). However, since the refined method produced a very similar result to the union, only the union is displayed in the figure. The important observation is that while posterior-decoding enables refined control over the recall/precision trade-off, the Viterbi decoding generates only three alignments, which cover only a small fraction of the curve at its high precision range. The union of Viterbi alignments achieves 0.531 recall at 0.913 precision, which is similar result to the 0.540 recall at 0.909 precision achieved using posterior-decoding with gap-factor and match-factor set to 1. However, unlike Viterbi, posterior-decoding produces alignments with much higher recall levels, by increasing the match-factor and decreasing the gap-factor. For example setting the gap-factor to 0.1 and match-factor to 1.2 results in alignments with 0.636 recall at 0.517 precision, and setting them to 0.05 and 1.5 results in 0.742 recall at 0.198 precision. Generally, the gap and match factor affect the accuracy of the alignments as expected. In particular, the alignments with the best AMA (0.889) and the best F_1-measure (0.678) are generated when the gap match factor are set to their natural values (1,1), which theoretically should maximize the expected AMA.

The performance of the pairwise alignments validates the underlying probabilistic model, showing it behaves as theoretically expected. However, the union of all pairwise alignments is not a valid MCA. To evaluate the MCA posterior decoding algorithm, we compared it to baseline MCAs. The baseline MCAs are constructed by using only the normalized-edit-distance $\frac{edit\_distance(c_k^i, c_l^j)}{\max\{|c_k^i|, |c_l^j|\}}$, and defining $c_k^i \leadsto^\delta c_l^j$ if and only if $normalized\_edit\_distance(c_k^i, c_l^j) \leq \delta$, where δ is a distance threshold. The final baseline MCA is constructed by taking the equivalence closure of all pairwise alignments, $\sim^\delta \triangleq \left(\leadsto^\delta \cup (\leadsto^\delta)^{-1}\right)^+$. The δ parameter can be used to control the recall/precision trade-off, since increasing it adds more position-pairs to the alignment, thus increasing recall, while decreasing it increases precision.

Figures 4 compares the performance of the CRF posterior-decoding MCAs with the baseline MCAs. The different MCAs were produced by varying the gap and match factors in the case of the posterior-decoding, and δ for the baseline MCAs. The CRF curve clearly dominates the baseline curve. However, they do overlap in range between 0.52 and 0.55 recall (0.84 and 0.90 precision). This is prob-

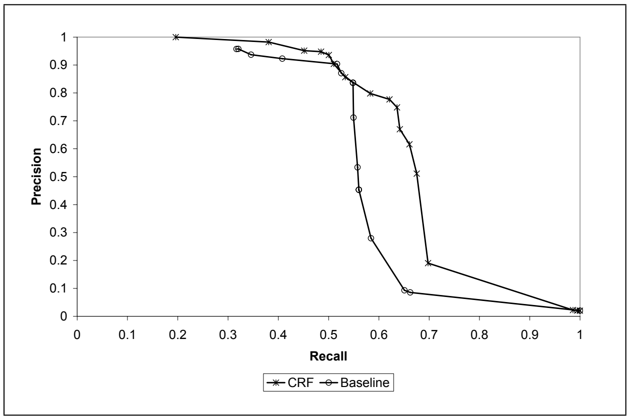


Figure 4: **Recall/Precision curve of MCAs comparing CRF with posterior decoding to normalized-edit-distance baseline.**

ably a range in which for this particular MCA the orthographic similarity is the most dominant feature. While the baseline curve drops sharply after that range, the posterior-decoding curve keeps improving recall up to 0.636 at 0.748 precision, before a major drop in precision. The additional recall is due to the ability of the CRF model to incorporate multiple overlapping features. In particular, the domain-specific features are important for aligning words and phrases that have little or no orthographic similarity. At the other end of the overlap range, the posterior-decoding achieves better precision than the baseline for the same recall levels. For example, the posterior decoding gets 0.381 recall at 0.982 precision compared with 0.346 at 0.937 for the baseline.

Unlike the pairwise alignment case, the neutral settings of the gap and match factors did not result in the best AMA score. This is due to the equivalence closure heuristic that results in MCAs that are too dense, since a single link between two equivalence classes causes them to merge. The best AMA score (0.886) is obtained by reducing the gap-factor to 0.5 and match-factor to 0.45, in order to compensate for the effect of the equivalence closure heuristic. For comparison, the best F_1-measure (0.690) is achieved by setting the gap and match factors to 0.75.

An error analysis on the latter MCA shows that out of 1400 unique errors, 1194 (85.3%) are false negatives (FN) and 206 (14.7%) false positives (FP). Most errors (more than 600) are due to misalignment of subtypes (e.g., *cdc*, *cdc6*, *cdc25A*), opposites (e.g., *phosphorylated* and *unphosphorylated*) and complex entities (e.g., *cell cycle* v.s. *cell line*). In addition, a large portion of FN errors are due to not aligning entities that belong to just four equivalence classes (e.g., 97 FN errors caused by terms in the class of *motif*, *site* and *domain*). Other types of errors include not aligning plural and singular forms of the same entities, aligning only part of multi-word entities, and incorrectly aligning orthographically similar entities that belong to different classes.

7 Conclusions

We have shown how to derive a posterior-decoding algorithm that aims at maximizing the expected utility for the MCA problem, as a substitute for the sequence-annealing algorithm for MSA. Adding a gap and match factor to the utility function allows control over the recall/precision trade-off when using posterior-decoding. Another advantage of optimizing the expected utility with posterior-decoding methods is the decoupling from the probabilistic model that generated the posterior probabilities. This allows the use of CRFs instead of HMMs with a similar posterior decoding algorithm.

Our experiments were limited by the size of the labeled data. However, the results support the theoretical predictions, and demonstrate the advantage of posterior-decoding over Viterbi decoding.

Since citances are still a relatively unexplored resource, it is still unclear whether the formulation we presented here for citance alignment is the most useful for applications that use citances for comparative analysis of bioscience text. Unlike biological sequence alignment, citance alignments are much more subjective, as they depend on a loose definition of semantic homology between entities. Even the definition of the basic entities can vary, since in many cases noun-compounds and other multi-word entities seem to be a more natural choice for basic elements of semantic homology and alignment. However, automatic segmentation and entity recognition are still difficult tasks in the bioscience text domain and so new methods are worth investigating.

Acknowledgements: We thank Phil Blunsom and Trevor Cohn for sharing their CRF-based word alignment code. This research was supported in part by NSF DBI 0317510.

References

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 65–72, Sydney, Australia, July. Association for Computational Linguistics.

Josias Braun-Blanquet. 1932. *Plant sociology: the study of plant communities*. McGraw-Hill, New York.

Francis Caillez and Pascale Kuntz. 1996. A contribution to the study of the metric and euclidean structures of dissimilarities. *Psychometrika*, 61(2):241–253.

Pedros Domingos. 2000. A unified bias-variance decomposition and its applications. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 231–238, Stanford, CA. Morgan Kaufmann.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis. Probablistic models of proteins and nucleic acids*. Cambridge University Press.

Jacob Falck, Niels Mailand, Randi G. Syljuasen, Jiri Bartek, and Jiri Lukas. 2001. The ATM-Chk2-Cdc25A checkpoint pathway guards against radioresistant DNA synthesis. *Nature*, 410(6830):842–847.

Eugene Garfield. 1955. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98. ACM Press.

Lynette Hirschman, Jong C. Park, Junichi Tsujii, Limsoon Wong, and Cathy H. Wu. 2002. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561.

Simon Lacoste-Julien, Ben Taskar, Dan Klein, and Michael I. Jordan. 2006. Word alignment via quadratic assignment. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 112–119, New York City, USA, June. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June. Association for Computational Linguistics.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Ben-Ami Lipetz. 1965. Improvements of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16:81–90.

Mengxiong Liu. 1993. Progress in documentation. the complexities of citation practice: A review of citation studies. *Journal of Documentation*, 49(4):370–408.

Evgeny Matusov, Richard Zens, and Hermann Ney. 2004. Symmetric word alignments for statistical machine translation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 219, Morristown, NJ, USA. Association for Computational Linguistics.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In Rada Mihalcea and Ted Pedersen, editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *HLT/EMNLP*, pages 81–88.

Preslav I. Nakov, Ariel S. Schwartz, and Marti A. Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *SIGIR'04 Workshop on Search and Discovery in Bioinformatics*.

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

Gian-Carlo Rota. 1964. The number of partitions of a set. *The American Mathematical Monthly*, 71(5):498–504, may.

Ralf Schlüter, Thomas Scharrenbach, Volker Steinbiss, and Hermann Ney. 2005. Bayes risk minimization using metric loss functions. In *Proceedings of the European Conference on Speech Communication and Technology, Interspeech*, pages 1449–1452, Portugal, September.

Ariel S. Schwartz and Lior Pachter. 2007. Multiple alignment by sequence annealing. *Bioinformatics*, 23(2):e24–29.

Ariel S. Schwartz, Eugene W. Myers, and Lior Pachter. 2006. Alignment metric accuracy. *arXiv:q-bio.QM/0510052*.

Ariel S. Schwartz. 2007. *Posterior Decoding Methods for Optimization and Accuracy Control of Multiple Alignments*. Ph.D. thesis, EECS Department, University of California, Berkeley.

Howard D. White. 2004. Citation analysis and discourse analysis revisited. *Applied Linguistics*, 25(1):89–116.