The applications of unsupervised learning to Japanese grapheme-phoneme alignment

Timothy Baldwin and Hozumi Tanaka
Tokyo Institute of Technology
{tim,tanaka}@cs.titech.ac.jp

Abstract
In this paper, we adapt the TF-IDF model to the Japanese grapheme-phoneme alignment task, by way of a simple statistical model and an incremental learning method. In the incremental learning method, grapheme-phoneme alignment paradigms are disambiguated one at a time according to the relative plausibility of the highest scoring alignment schema, and the statistical model is re-trained accordingly. On limited evaluation, the learning method achieved an accuracy of 93.28%, representing a slight improvement over a baseline rule-based method.

1 Introduction
The objective of this paper is to analyse the applicability of statistical and learning methods to automated grapheme-phoneme alignment in Japanese, without reliance on pre-annotated training data or any form of supervision. The two principal models proposed herein are a simple statistical model non-reliant on learning techniques, and an incremental learning method deriving therefrom, incorporating automated “pseudo-supervision” drawing on prior alignments. The incremental learning method selects a single alignment candidate to accept at each iteration, and adjusts the statistical model accordingly to aid in the subsequent disambiguation of residue G-P tuples.

Grapheme-phoneme (“G-P”) alignment is defined as the task of maximally segmenting a grapheme compound into mora-phonemic units, and aligning each unit to the corresponding substring in the phonemic compound (Bilac et al., 1990). Its main use is in portrayal of the phonological interaction between adjoining grapheme segments, and also implicit description of the range of readings each grapheme segment can take. We further suggest that a large-scale database of maximally aligned G-P tuples has applications within the more conventional task of G-P translation (Klatt, 1987; Huang et al., 1994; Divay and Vitale, 1997).

Our particular interest in developing a database of G-P tuples is to apply it in the development of a kanji tester which can dynamically predict plausibly incorrect readings for a given grapheme string. For this purpose, we require as great a coverage of grapheme strings as possible, and the proposed system has thus been designed to exhaustively align the input set of G-P tuples, sacrificing precision for 100% recall.

‘Grapheme string’ in this research refers to the maximal kanji representation of a given word or compound, and ‘phoneme string’ refers to the kana (hiragana and/or katakana) mora correlate.1 By ‘maximal' segmentation is meant that the grapheme string must be segmented to the degree that each segment corresponds to a self-contained component of the phonemic description of that compound, and that no segment can be further segmented into aligning sub-segments. The statement of ‘maximality’ of segmentation is qualified by the condition that each segment must constitute a morpho-phonemic unit, in that for conjugating parts-of-speech, namely verbs and adjectives, the conjugating suffix must be contained in the same segment as the stem.

By way of illustration of the alignment process, let us consider the example of the verb ka-n-sya-su-ru [ka-nsya-su-ru] “to thank/be thankful”,2 a portion of the 35 member alignment paradigm for which is given in Fig. 1. The importance of maximality of alignment is observable by way of align35, which constitutes a legal (under-)alignment of the correct solution in align1. Here, there is scope for further segmentation, as evidenced by the replaceability of に by its phonemic content of な in isolation of
1Our description of kana as phoneme units represents a slight abuse of terminology, in that individual kana characters are uniquely associated with a broad phonetic transcription potentially extending over multiple phones. Note, however, that in abstracting away from this meta-phonemic representation, we are freed from consideration of low-level phonological concerns such as phoneme connection constraints.
2So as to make this paper as accessible as possible to readers not familiar with Japanese, hiragana and katakana characters have been transliterated into Latin script throughout this paper and are essentially treated as being identical. The graphemic kanji character set has been preserved in its original form to give the reader a feel for the significance of the kana-kanji dichotomy. For both the grapheme and phoneme strings, character boundaries are indicated by “-“ and segment boundaries (which double as character boundaries) indicated by “®”.

...
2 The grapheme-phoneme alignment process

Grapheme-phoneme alignment is performed as a four-stage process: (a) detection of lexical alternations and removal of lexical alternates from the input, (b) determination of all possible G-P alignment schemas, (c) pruning of alignments through phonological constraints, and (d) scoring of all final candidate alignments, and determination of the final solution accordingly.

2.1 Lexical alternation

Lexical alternation is defined as the condition of there being multiple lexical spell-outs for a given phonetic content, all sharing the same basic semantics and kanji component. For Japanese, this can arise as a result of the replaceability of kanji and their corresponding kana (i.e. maze-gaki, as seen above for ka-n-sya-su-ru), or alternatively for okurigana. Okurigana comprise a (generally) inflecting kana suffix to a kanji stem, where the combination of the kanji stem and okurigana form a single morphophonemic segment; an example of okurigana is seen for the ru of 送-ru [o-ku-ru] "to send", with inflects to re in the imperative, for example. Okurigana-based lexical alternation occurs when phonetic content is conflated with or prised apart from the stem kanji, by way of okurigana optionality. An example of this occurs for the verb か-わ-る "to change", lexicalisable either as ～2-ル or ～.-ワ-ル, with the underlined wa conflating with the kanji stem of 異 in the former (basic) case for the same phonetic content. Note that okurigana never occur as alternating prefixes to kanji.

Detection of okurigana alternates is achieved by way of analysing the graphemic form of G-P tuples sharing the same phonetic content, and aligning the graphemic component of each such corresponding tuple to determine kanji correspondence. All instances of okurigana-based lexical alternation are clustered together, and alternates of the 'basic' form removed from input. The basic form is defined as that with maximal phonemic conflation, that is minimal kana content in the grapheme string. In this way, we can: (a) enforce consistency of analysis for all okurigana alternates, (b) apply alignment constraints across the full set of lexical alternates, and (c) avoid having multiple realisations of the same
Intra-syllabic segments cannot exist for kana

Graphemic kana must align with a direct kana

A demarkation in script form indicates a segmentation ("e"), and/or grapheme-phoneme alignment to grapheme segmentation ("G"), phoneme segmentation ("P"), respectively:

\( p_1 \) A demarkation in script form indicates a segmentation boundary, except for the case of kanji-hiragana boundaries. [G]

\( p_2 \) Graphemic kana must align with a direct kana equivalent in the phoneme string. [G-P]

\( p_3 \) Intra-syllabic segments cannot exist for kana strings [G-P]

The length of a kanji substring must be equal to or less than the syllable length of the corresponding phoneme substring. [G-P]

Constraint \( p_1 \) produces the result that a segment boundary must exist at every changeover between hiragana and katakana, or kanji and katakana, and from hiragana to kanji. The exceptional treatment of kanji-hiragana changeovers is designed to facilitate the recognition of full verb and adjective morphophonemic units, as these two parts-of-speech involve conjugating kana suffices and also the potential for furigana-based lexical alternation. Note that for \( \text{align}_n \) in Fig. 1, we do in fact have a segment boundary at the kanji-hiragana changeover ???.

Constraint \( p_2 \) polishes the essentially phonemic nature of kana, in disallowing alignment of kana segments of non-corresponding phonetic content. In the case of Fig. 1, \( p_2 \) would lead to the disallowance of \( \text{align}_n \) due to the alignment of \((...®sya-su-ru) — (...®sya-su-ru)\). Here, we would disallow a segment boundary to exist between ka and n, and as such prune off \( \text{align}_n \) in Fig. 1.

Finally, \( p_4 \) requires that each kanji character leads to a phoneme substring at least one syllable in length, irrespective of whether that single kanji comprises the head of a morpho-phonemic unit or combines with adjoined kanji to form a multiple-phoneme segment. A two kanji segment is required, therefore, to align with a phoneme substring at least two syllables in length. "kinese" could thus not align with the mono-syllabic ka-n, leading once again to the pruning of \( \text{align}_n \).

Note, there also exists scope to apply intra-segmental phonological constraints such as Lyman’s Law (Itô and Mester, 1995, p. 819), which is left as an item for future research.

3 Scoring method

The scoring method utilised in this research for both method-1 and method-2 is an adaptation of the TF-IDF model (Salton and Buckley, 1990), best known in the context of term weighting for information retrieval (“IR”) tasks. The main differences between our usage of the TF-IDF model and standard usage within IR circles, come in the counting of frequencies (method-1 and method-2) and the incremental updating of the statistical model/weighting of terms according to system “conviction” (method-2).

That we should require a special means of counting frequencies is a direct consequence of the two proposed methods dynamically determining segmentation schemas as a component of the alignment process. We integrate the segmentation and alignment processes by taking the frequency of occurrence of a given segment as the number of G-P tuples for which
that segment is contained in the alignment paradigm in an identical lexical context.

By adopting this approach of alignment potential-based frequency, we do not discount the possibility of any alignment licenced by the constraints given above, but at the same time are unable to commit ourselves to any alignment schema we believe is correct. In method-2, therefore, we combine the existential-based statistical modelling of method-1 for non-disambiguated alignment paradigms (ψ in Fig. 2), with a means of dynamically updating the statistical model based on selectively disambiguated alignment paradigms (ω in Fig. 2).

Alignment paradigms are selected for disambiguation based on the degree of discrimination between the top- and second-ranking alignment schemas, and term frequencies found in solution alignments in ω weighted above those found in the alignment paradigms of ψ. Note that by disambiguating a particular alignment paradigm, we are both identifying that alignment schema we believe to be correct, and disallowing all alternate alignments. As such, updating of the statistical model reflects on all terms contained in rejected alignment schemas. This results in a rescoring of all alignments containing affected terms.

3.1 Why tf-idf?

The applicability of the TF-IDF model to G-P alignment can be understood intuitively by considering each grapheme segment type as a document, the associated phonemic segments across all G-P tuples as terms, and the left and right graphemic/phonemic contexts of the current grapheme/phoneme strings, as the document context.

The TF-IDF model maximally weights terms which occur frequently within a given document (TF) but relatively infrequently within other documents (IDF). For G-P alignment, we maximally weight readings (aligned phoneme strings) which co-occur frequently with a given grapheme string, but are observed infrequently in the given lexical context. That is, we score up terms which occur with high relative frequency and maximum diversity of lexical context, and score down terms which either occur infrequently or occur only in restricted lexical contexts. In this way, we are able to penalise under-alignment by way of a diminished TF score (as the given over-alignment will be reproducible for only a small component of instances of either the same grapheme or phoneme string). By calculating individual TF-IDF scores for each each aligned segment and combining them to produce a single overall score for the alignment, we are able to balance up selection of the optimal overall alignment for the tuple.

A subtle advantage in using the TF-IDF model in the manner proposed above is that it has no sense of "appropriate" segment size. While single characters provide a lower bound on segment size and the full string in question provides a dynamic upper bound, our only constraint within these bounds is that segment size must follow character boundaries. In the given context of Japanese G-P alignment, it commonly occurs that both phoneme and grapheme segments extend over multiple characters (for the 5000 member test data used for evaluation purposes, the average phoneme and grapheme segment sizes were 1.93 and 1.20 characters, respectively). Indeed, despite the general perception of grapheme segments as containing a single kanji, multiple kanji were found in grapheme segments for 0.9% of G-P tuples in the test data (see below), including instances of the type 昨日 [ki-n] “yesterday” and 茄子 [na-su] “eggplant”. The TF-IDF model can handle such examples because of the scarcity of alignment candidates sharing any of the unit-kanji readings produced through segmentation of such grapheme strings. That is, we would not expect to locate the partial alignment {...○ impose...}-(...○sw○...), for example, with significant frequency in the remainder of the alignment data, whereas we may find the partial alignment {...○茄-子○...}-(...○na-sw○... elsewhere. Even if there were only one instance of this alignment type in the system data, the combination of the diminished scores for {...○茄○...}-(...○na○...) and {...○子○...}-(...○sw○... would lead to an overall TF-IDF score for the associated segmentation well below the TF-based score for the full string-based alignment (see below).

3.2 Counting frequencies

To be able to apply the basis of the TF-IDF model, we first need to have means of calculating term frequencies. Given that both methods are designed to operate independently of annotated training data, we have no means of bootstrapping the system.3

3Not strictly true, as there are a significant number of G-P tuples where the alignment constraints produce full dis-
Term frequencies are thus defined to be an indication of the number of G-P tuples for which the full alignment paradigm contains the given term, without consideration of whether that instance occurs within a correct alignment or not. This can be represented as in equation (1), in the case of $freq((g,p))$, where $p$ is the phoneme string aligning with grapheme string $g$ and $phon\_var(p)$ describes the set of phonological alternates of $p$.

Phonological alternates are predictable instances of phonological alternation from a base form $p$, with the most widespread types of phonological alternation being "sequential voicing" (Tsujimura, 1996, 54-63) and gemination; if no method were provided to cluster frequencies for phonological alternates together, data sparseness and skewing of the statistical model would inevitably result. The current system has no way of predicting exactly what form of phonological alternation is likely to occur in what lexical context. One observation which can be made, however, is that phonological alternation affects only the phoneme string, and occurs only at the interface between adjacent phoneme segments on a single syllable level. It is thus possible to establish phonological equivalence classes at the unit syllable level, and use these to determine the maximum scope of phonological alternation which could realistically be expected of a given phoneme string.

Formally, for a given phoneme string $p = s_1s_2...s_n$ aligning with grapheme string $g$, where each $s_i$ is a syllable unit, we thus generate a regular expression of all plausible phonological alternations $\{s_a|s_b|...\}s_2...\{s_a|s_b|...\}$, where $\{s_a|s_b|...\}$ and $\{s_a|s_b|...\}$ are the phonological equivalence classes for $s_1$ and $s_n$ respectively. For example, given the phoneme string $ka$-$ka$-$ku$-$ku$, we would generate the string-level equivalence class $\{ka|ga\} \{ku|gu\}$, where the $ka$-$ga$ and $ku$-$gu$ unit grapheme alternations are attributable to sequential voicing, and the $ku$-$gu$ alternation to gemination.

The frequencies of all phonological alternations subsumed by the string-level equivalence class are then combined within $freq((g,p))$. We are able to handle phonological alternation within the confines of the original statistical formulation by virtue of the fact that the grapheme string is unchanged under phonological alternation, and as such the combined frequencies of alternates can never exceed the frequency of the associated grapheme string segment. This guarantees a $tf$ value in the range $[0,1]$.

### 3.3 The modified tf-idf model

Our interpretation of the TF-IDF model is given in equation (2), where $g$ is a grapheme unit, $p$ a phoneme unit and $cttz$ some lexical context for $(g,p)$ within the current alignment; $freq(g)$, $freq((g,p))$ and $freq((g,p,cttz))$ are the frequencies of occurrence of $g$, the tuple $(g,p)$, and the tuple $(g,p,cttz)$ in lexical context $cttz$, respectively. The subtractions by a factor of one are designed to remove from calculation the single occurrences of $(g,p)$ and $(g,p,cttz)$ in the current alignment, and $\alpha$ is an additive smoothing constant, where $0 < \alpha < 1$.

Consideration of lexical context for a given tuple $(g,p)$ is four-fold, made up of the single character immediately adjacent to $g$ in the grapheme string and single syllable immediately adjacent to $p$ in the phoneme string, for both the left and right directions. In the case that $(g,p)$ is a prefix of the overall G-P string pair, we disregard left lexical context and simply score according to $tf$, that is the ratio of occurrence of $g$ with reading $p$, for the two left context scores. Correspondingly in the case of $(g,p)$ being a suffix, we disregard right context. The four resultant scores are then combined by taking the arithmetic mean of the averaged scores for each segment pairing, with the exception of full kana-based grapheme segments which are removed from computation altogether.

### 3.4 Verb/adjective conjugation

There is one remaining form of commonly-occurring alternation which cannot be resolved easily within the confines of the TF-IDF model. This is verbal/adjectival conjugation, and is difficult to cope with given the existing statistical formulation because it occurs concurrently at both the grapheme and phoneme levels (i.e. we have no immediate ceiling on combined frequencies as was the case for phonological alternation). We model conjugation-based alternation by postulating verb paradigms based on conjugational analysis of the kana suffix to a given stem (Baldwin, 1998). This postulation of verb paradigms is performed independently of any static verb dictionary, and is achieved simply by clustering legal verb stem-inflectional suffix segments according to verb stem and conjugational class. For example, for the aligned segment $\langle$解$\rangle$ $\langle$-く$\rangle$ (which constitutes the non-past form of the verb $tok$-$u$ "to undo"), conjugational analysis would reveal the possibility of the segment being comprised of the verb stem $tok$ and inflectional suffix of $u$. Subsequent analysis of the corpus may well unearth what constitute conjugates of the same verb postulate, in $to$-$ki$, for example. This could then be complemented by consideration of phonological alternation as above, to produce the verb paradigm $(toku, doku, toki, dokis)$. To be able to combine scoring of verb conjugates of the same verb paradigm within the original formulation (i.e. TF), we now require some base form of the verb which is guaranteed to occur with at least the same frequency as all its alternates, and hence constrain the value of TF to the range $[0,1]$.

For method-1, it is possible to consider the (invariant) verb stem as the base form of the verb. In equation (2), we thus replace $freq(g)$ by $freq_{-1}(g)$, that is the frequency of the graphemic component of verb stem $g$ (irrespective of whether

---

4 Here, $\phi$ designates the head of a long consonant, also indicated by $/Q/$ in phonological theory.

5 Although discussion here refers exclusively to verbs, (conjugating) adjectives are handled in exactly the same manner.
or not it is contained within a recognised conjugation of the verb, and also irrespective of what phoneme segment it aligns with), and in equation (1), \( \text{phon}_{\text{var}}(p) \) becomes the augmented set of all phonological alternates of all conjugations of the verb \( p \). Scoring is now carried out by way of the simple TF model, without recourse to IDF. This decision was made based on the observation that inherent delimitation of verb conjugates is provided through inflection-based analysis, such that there is little danger of under- or over-aligning the segment in question.

This leaves us in the position of having two separate means of scoring verb conjugate postulates, one via the basic TF-IDF formulation described in Section 3.3, and one through the TF-based conjugation model described in the above paragraph. In cases of such analytical ambiguity, there is potential for the model described in the above paragraph. In cases of little danger of under- or over-aligning the segment in question, and do not commit ourselves a priori to either analysis.

This completes the formulation of method-1.6

In method-2, on the other hand, we are unable to find our frequency count on the base form of the verb, as the whole verb conjugate constitutes a single morpho-phonemic segment for disambiguated alignments. As such, no instance of the verb stem can be found as an individual segment. We thus modify our definition of \( \text{freq}_{\text{var}}(g) \) somewhat to \( \text{freq}_{\text{var}}(g) \): the frequency of all G-P tuples for which there is an alignment candidate containing a conjugate existing in the same inflection paradigm as \( g \). While this provides us with a ceiling for the raw frequencies of verbs and adjectives, weighting up of verb conjugates is achieved through the use of two weighting segments is always in the range \([0, 1]\). In practice, this means that the score for a given verb inflection is initialised to \( \text{cuf}_1 \) and tends to converge to either 0 (in the case of the postulated verb paradigm being rejected for each conjugate instance), or 1 (in the case of it being accepted).

### 3.5 Incrementally learning with method-2

We are now in the position of being able to set method-2 running, and the only remaining consideration is exactly how we should select which alignment paradigm to disambiguate at each iteration, and how to implement the incrementality of the learning method.

Selection of the alignment paradigm for disambiguation is achieved through the application of a discriminative metric. Two metrics were tentatively trialled for this purpose. The first consists of the simple ratio \( \frac{\text{freq}_{\text{var}}(g)}{\text{freq}_{\text{var}}(g) + \text{cuf}_1} \) between the highest and second highest ranking scores \( s_1 \) and \( s_2 \) ("the odds ratio"), in the manner of (Dagan and Itai, 1994). The second discriminative metric \( (\text{dm}_2) \) is a slight variation on this whereby we take the log of the ratio of the highest ranking score to the second ranking score ("the log odds ratio"), and multiply it by the highest ranking score, i.e. \( s_1 \log \frac{s_1}{s_2} \). The G-P tuples contained in \( \psi \) are ranked in descending order according to the particular discriminative metric of use, and the G-P tuple with the highest rank (i.e. with greatest system "conviction" in the top-ranking alignment candidate) is disambiguated based on the top-scoring alignment candidate.

The first discriminative metric is heuristic, and based on the intuition that we are after maximum disparity in score between the first and second ranked candidates. The second discriminative metric, on the other hand, is designed to balance up maximisation of both \( s_1 \) and the relative disparity between \( s_1 \) and \( s_2 \). Note that, unlike Dagan and Itai (1994), we give no consideration to statistical confidence as we are after 100% recall, whatever the cost to precision.

To this point, the only difference over method-1 is the sequence in which solutions are output. However, by singling out a G-P alignment candidate of maximum discrimination on each iteration, it now becomes possible to refine the statistical model by training it on aligned output (i.e. G-P tuples stored in \( \omega \) in Fig. 2), hence: (a) alleviating statistics deriving from less-plausible alignments, and (b) weighting up term frequencies found in final disambiguated alignments. Neither of these processes are possible under the simple statistical model as all alignments are processed in parallel, and the system is unable to commit itself to the plausibility of any given alignment in scoring others.

The weighting up of terms found in solution alignments is achieved through the use of two weighting factors on term frequencies, one for terms found in candidate alignments (\( \psi \)) and one for terms found in solution alignments (\( \omega \)), namely the candidate weighting factor (cuf) and solution weighting factor (suf), respectively; naturally, \( 0 < \alpha < \text{cuf} \leq \text{suf} \).

### 4 Evaluation

As a test set, a set of 5000 G-P tuples was randomly extracted from the EDICT English-Japanese dictionary7 and Shinmeikai Japanese dictionary (Nagawa, 1981) and each tuple annotated with its alignment for evaluation purposes. So as to be able to properly evaluate the success of application of the alignment constraints, we further augmented the original 5000 G-P tuples with 1403 lexical alternates thereof (so as to provide full scope for constraint-based pruning). Our motivation in using this limited data set was to be able to run method-2 to completion and attain empirically comparable results for the two proposed methods.

---

6For discussion of further variations on method-1, see (Baldwin and Tanaka, 1999).

7ftp://ftp.cc.monash.edu.au/pub/nihongo
In evaluation, method-1 was used with the α smoothing constant set variably to \{0.25, 0.05, 0.001, 0.0001\}. For method-2, \textit{cuf} and \textit{suf} were fixed at 0.5 and 1.0 respectively, and α set variably to \{0.05, 0.0001\} for discriminative metric \(dm_1\), and \{0.25, 0.05, 0.001\} for \(dm_2\).

By way of a baseline for evaluation, we used the rule-based method proposed by Bilac et al. (1999), which achieved an alignment accuracy of 92.90% when run over the full dictionary file of 59744 entries and empirically evaluated on the same 5000-tuple data set as was used for method-1 and method-2. Note that the Bilac system requires a training set of standard readings for each unit kanji and also a verb conjugational dictionary, whereas both our proposed methods have no reliance on external evidence. It is also worth emphasising that our methods were heavily handicapped over the rule-based method, in that they were not able to apply statistics derived from the remaining 52744 entries in refining their respective statistical models. However, in terms of empirical evaluation of the three methods, the respective system accuracies are directly comparable.

As evidenced in Fig. 3, method-1 achieved a maximum accuracy of 86.74% (with \(\alpha = 0.0001\)), significantly below that of the baseline method. Based on the curve for method-1, it would appear that the method performs best with infinitesimally small \(\alpha\) values. This perhaps points to limitations in our "plus constant \(\alpha\)" smoothing methodology. In stark contrast, method-2, achieved a maximum accuracy of 93.28% (using \(dm_2\), with \(\alpha = 0.05\), just outstripping the baseline method despite its handicap in terms of diversity of input data. Little difference was seen between accuracies for discriminative metrics \(dm_1\) and \(dm_2\), although \(dm_2\) generally performed marginally better. For the given \textit{cuf} and \textit{tuf} values, it would appear that an \(\alpha\) value around 0.05 is optimal, providing an interesting comparison with the seemingly asymptotic nature of the method-1 curve. While we are unable to present the results here, varying the relative values of \textit{cuf} and \textit{tuf} produced little difference over the accuracies in Fig. 3, for comparative \(\alpha\) values.

The most common type of system error for method-1 was under-alignment (where the correct alignment is properly subsumed by the system alignment). That the system accuracy increases with diminishing \(\alpha\) value is a result of decreases in under-alignment outweighing increases in over-alignment and over-segmentation on conjugating morphemes. For method-2, the greatest single error type is over-segmentation of conjugating morphemes (principally verbs), accounting for 58.95% of all errors for \(dm_2\) with \(\alpha\) set to 0.001. It would appear that for relatively larger values of \(\alpha\), instances of under-alignment increase, and for relatively smaller values of \(\alpha\), instances of over-alignment and over-segmentation increase.

So as to get an insight into its true potential, we re-did evaluation of method-1, over the full dictionary set this time with \(\alpha\) set to 0.05 (using the same 5000 tuples for evaluation as before). This produced an accuracy of 93.96%, pointing to the potential for a even higher accuracy for method-2 over the full dictionary set.

Analysis of the effectiveness of the lexical and phonological constraints indicated that we are able to reduce the cardinality of alignment by almost 75%, from 13.80 to 4.10, on average. Indeed, full disambiguation was possible for 603 of the 5000 entries (including 480 singleton entries). Importantly, there were no instances of the correct alignment being pruned due to over-constraint. The individual constraints were activated with the frequencies indicated below, with constraints higher in the table taking precedence over those lower in the table in the case of a given alignment violating more than one constraint.

| Times activated | Relative freq. of application |
|-----------------|-------------------------------|
| \(l\)           | 18481                         | 34.41% |
| \(p_1\)         | 9076                          | 16.90% |
| \(p_2\)         | 11383                         | 21.19% |
| \(p_3\)         | 9292                          | 17.30% |
| \(p_4\)         | 14297                         | 26.62% |

To further examine the correspondence between the size of the discriminative ratio and system accuracy for method-2, we plotted both the system accuracy and discriminative value against the rank of sys-
tem output (Fig. 4 - based on $dm_2$ with $\alpha = 0.05$). Here, we disregard all alignments where constraints produced full disambiguation (603 instances), such that the rank of the first statistically disambiguated input is 604. The indicated accuracies and discriminative values are averaged over discrete corridors of \( \approx 220 \) entries centering on the given output ranks. Looking to the results, it is important firstly to notice that we realise an accuracy of 100% in the initial stages of output (up to rank 1703), which progressively degrades down to 92.38% over the final corridor with zero discriminative. Note also that whereas the discriminative curve is monotonically decreasing when averaged over the given corridor, in practice local maximums do exist, attributable to the situation where re-training of the statistical model produces inflation of the maximum discriminative value.

5 Other applications of this research

Other than the constraints described in Section 2 and frequency determination techniques, the proposed methodology is theoretically scalable to any domain where two streams of chunked information require alignment. This suggests applications to the extraction of translation pairs from aligned bilingual corpora (Gale and Church, 1991; Kupiec, 1993; Smadja et al., 1996), where the system input would be made up of aligned strings (generally sentences) in the two languages. Given that we can devise some way of creating an alignment paradigm between the two input segments, it is possible to apply the scoring and learning methods proposed herein in their existing forms. Note, however, that in the case of translation pair extraction, there is a real possibility of the alignment mapping being many-to-many, and crossing over of alignment is expected to occur readily. In fact, it may occur that there is a residue of unaligned segments in either or both languages, as could easily occur if one language included zero anaphora. It may, therefore, be desirable to apply a dynamic threshold on the discriminative ratio (cf. (Dagan and Itai, 1994)) to accept only those translation pairs with sufficiently high statistical confidence, for example.

6 Conclusion

In this paper, we proposed an adaptation of the TF-IDF model to Japanese grapheme-phoneme alignment. We then went on to extend the basic statistical method to devise a fully unsupervised learning method, by way of a two discrimination-based metrics and incremental refinement of the statistical model. Experimentation suggested that the proposed learning method marginally outperforms both a baseline rule-based method and the non-incremental statistical method.

Items of future research include expanding evaluation of the incremental learning method to the full dictionary file used in this research, as well as to other Japanese dictionaries/genres and other languages.

Acknowledgements

The authors would like to thank Assoc. Prof. Noguchi, Assoc. Prof. Tokunaga, Masahiro Ueki, Christoph Neumann and two anonymous reviewers for their insightful comments on earlier versions of this paper. We also pay tribute to the heroic efforts of Slaven Bilac in implementing the rule-based version of the system.

References

T. Baldwin and H. Tanaka. 1999. Automated Japanese grapheme-phoneme alignment. In Proceedings of the International Conference on Cognitive Science, Tokyo. (to appear).

T. Baldwin. 1998. The Analysis of Japanese Relative Clauses. Master's thesis, Tokyo Institute of Technology.

S. Bilac, T. Baldwin, and H. Tanaka. 1999. Incremental Japanese grapheme-phoneme alignment. In Information Processing Society of Japan SIG Notes, volume 99-NL-209, pages 47–54.

I. Dagan and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. Computational Linguistics, 20(4):563–96.

M. Divay and A.J. Vitale. 1997. Algorithms for grapheme-phoneme translation for English and French: Applications for database searches and speech synthesis. Computational Linguistics, 23(4):495–523.

W.A. Gale and K.W. Church. 1991. Identifying word correspondences in parallel texts. In Proceedings of the Fourth DARPA Speech and Natural Language Workshop, pages 152–7. Morgan Kaufmann.

C.B. Huang, M.A. Son-Bell, and D.M. Baggett. 1994. Generation of pronunciations from orthographies using transformation-based error-driven learning. In Proc. of the International Conference on Speech and Language Processing, pages 411–4.

J. Itô and R. Armin Mester. 1995. Japanese phonology. In J.A. Goldsmith, editor, The Handbook of Phonological Theory, chapter 29, pages 817–38. Blackwell.

D.H. Klatt. 1987. Review of text to speech conversion for English. Journal of the Acoustic Society of America, 82(3):737–793.

J. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In Proceedings of the 31st Annual Meeting of the ACL, pages 17–22.

K. Nagasawa, editor. 1981. Shinmeikai Dictionary. Sanseido Publishers.

G. Salton and C. Buckley. 1990. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4):288–97.

F. Smadja, K.R. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics, 22(1):1–38.

N. Tsujimura. 1996. An Introduction to Japanese Linguistics. Blackwell.