A COMPARISON OF CORPUS-BASED TECHNIQUES FOR RESTORING ACCENTS IN SPANISH AND FRENCH TEXT

DAVID YAROWSKY
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218-2694

Abstract. This chapter will explore and compare three corpus-based techniques for lexical ambiguity resolution, focusing on the problem of restoring missing accents to Spanish and French text. Many of the ambiguities created by missing accents are differences in part of speech: hence one of the methods considered is an N-gram tagger using Viterbi decoding, such as is found in stochastic part-of-speech taggers. A second technique, Bayesian classification, has been successfully applied to word-sense disambiguation and is well suited for some of the semantic ambiguities which arise from missing accents. The third approach, based on decision lists, combines the strengths of the two other methods, incorporating both local syntactic patterns and more distant collocational evidence, and outperforms them both. The problem of accent restoration is particularly well suited for demonstrating and testing the capabilities of the given algorithms because it requires the resolution of both semantic and syntactic ambiguity, and offers an objective ground truth for automatic evaluation. The problem is also a practical one with immediate application.

1. Problem Description

Accent restoration is closely related to several lexical disambiguation problems. It involves aspects of both word-sense disambiguation and part-of-speech tagging. While not as widely cited as these other tasks, it nonetheless offers considerable benefits as a case study, and is particularly useful for evaluating and comparing the disambiguation algorithms considered here. Specifically:

- It requires the resolution of both syntactic and semantic ambiguities, and is representative of many of the issues that arise in several important types of lexical ambiguity resolution.
- Unlike many ambiguity resolution tasks which depend on human annotations or judgements for evaluation, this problem supports fully automatic evaluation and an innate, plentiful and objective ground truth: text with accents may be artificially stripped, leaving accentless text for testing purposes with a known gold standard for evaluation.
- The problem has immediate and practical application, both as a standalone product and a front-end component to multilingual NLP systems. There is also a potentially large commercial market in its use in grammar and spelling correctors, and in aids for inserting the proper diacritics automatically during on-line typing. Such a tool would be particularly useful for typing Spanish or French on Anglo-centric computer keyboards, where entering accents and other diacritic marks every few keystrokes can be laborious.

Thus while accent restoration may not be be the prototypical member of the class of lexical-ambiguity resolution problems, it is an especially useful one for describing, evaluating, and comparing proposed solutions to this class of problems.

Accent ambiguities arise routinely under a number of circumstances in Spanish and French[1]. It is traditional in both languages for diacritics to be omitted from capitalized letters. This is particularly a problem in all-capitalized text such as headlines. Accents in on-line text may also be systematically stripped by many computational processes which are not 8-bit clean (such as some e-mail transmissions), and may be frequently omitted by Spanish and French typists in informal computer correspondence.

Limited space precludes a full discussion of the range of accent pattern ambiguities encountered; see Yarowsky (1994) for more detail. Discussion here will focus on the following types of ambiguity in Spanish: The most common ambiguity is between the endings *-o* and *-ó*, as in *marco* vs. *marcó*. This is typically both a verb-tense and part-of-speech ambiguity. The second most common general ambiguity is between the past-subjunctive and future tenses of nearly all -ar verbs (eg: *terminara* vs. *terminará*), both of which are 3rd person singular forms. This is a particularly challenging class and is not readily amenable to traditional part-of-speech tagging algo-

[1]For brevity, the term *accent* will typically refer to the general class of accents and other diacritics, including ê,è,é,ö etc. The term *accent restoration* should more accurately be called *diacritic restoration*.

rithms such as local trigram-based taggers. Other ambiguities include function words (*mi* vs.*mí*) and purely semantic ambiguities such as *secretaria* (secretary) vs. *secretaría* (secretariat). The distribution of ambiguity types in French is similar, including the frequent part-of-speech and/or tense ambiguity between *-e* and *-é* endings, and numerous semantic ambiguities such as *traité/traite* (treaty/draft).

2. Comparison of Algorithms

This section will describe the application of 4 algorithms to the problem of accent restoration, outlining the details of the implementations and performance achieved.

2.1. METHOD 1: BASELINE

The vast majority of tokens in Spanish and French exhibit only one accent pattern. And in cases where there is ambiguity, one pattern is typically dominant. Thus one can achieve surprisingly good performance by using only the most common accent pattern for each token. This baseline approach is the standard by which all other techniques will be measured.

Initial corpus analysis will yield accent pattern distributions such as in Table 1 (for French):

TABLE 1. Sample accent pattern distributions for French

De-accented Form	Accent Pattern	%	Number
cesse	cesse	53%	669
	cessé	47%	593
cout	coût	100%	330
couta	coûta	100%	41
coute	coûté	53%	107
	coûte	47%	96
cote	côté	69%	2645
	côte	28%	1040
	cote	3%	99
	coté	<1%	15
cotiere	côtière	100%	296

Measured over all tokens, the baseline approach achieves 98.7% mean accuracy for Spanish and 97.6% mean accuracy for French. A breakdown of the baseline performance on the words used in the comparative study is given in Table 2, under the column labeled "BaseL". While this base performance may seem high, it still produces an error every 40-75 words in text. More importantly, the cases that it misses are precisely those where accents resolve an ambiguity, and are thus the most important to handle correctly. Some attempt to resolve these ambiguities is clearly warranted.

2.2. METHOD 2: N-GRAM TAGGER

Since most of the ambiguities due to missing accents correspond to differences in parts-of-speech, it is natural to consider the algorithm most commonly applied to the problem of part-of-speech tagging, namely the Markov or N-Gram tagger. This approach was first widely publicized in (Church, 1988) and has become the standard in the field[2].

It is not necessary, however, to train a full part-of-speech tagger for Spanish and French to restore accents. Many part-of-speech distinctions have no direct bearing on choice of accent pattern. It may be advantageous to build an n-gram tagger which focuses only on the distinctions necessary to resolve the major accent ambiguities (e.g., *-o/-ó*, *-ara/-ará*, *-aran/-arán* in Spanish), ideally using only information available in an unannotated corpus.

One natural approach in morphologically rich languages such as Spanish and French is to build a model not of part-of-speech sequences, but of *suffix* sequences. It would be desirable if the patterns of suffixes and a small set of function words in nearby context were adequate to disambiguate ambiguous forms. The only linguistic knowledge that would be necessary then is a list of suffixes and function words in the language. Given such information, it is straight-forward to create a training set such as the following, with words annotated with their suffix or function word label, and the word form then stripped of any accents:

la posición anunció oficialmente que $\Longrightarrow$

```
... la/LA posicion/-IÓN anuncio/-Ó oficialmente/-MENTE que/QUE ...
```

cambiar el anuncio utilizado ... a mí $\Longrightarrow$

```
... cambiar/-AR el/EL anuncio/-O utilizado/-ADO ... a/A mi/MÍ ...
```

[2]Note that because the techniques in Methods 2 and 3 have been so thoroughly presented elsewhere, they will be covered somewhat briefly here to allow more space to be devoted to the new Method 4.

Using an N-Gram tagger trained on such data, one can recover the most probable suffix annotations for a sequence of de-accented text. Because a word's accent pattern is almost always unambiguous given a suffix and can be described in a table such as in Method 1 above, the disambiguation process is a straightforward application of the channel model. The actual algorithm used in this experiment is described in (Rabiner 1989) and (Paul 1990). The B-matrix emit probabilities are defined as $B[TAG, deaccented\_token] \equiv p(deaccented\_token|TAG)$, with transition probabilities defined analogously. The most probable tag sequence for new test data is recovered using a standard Viterbi decoder, implemented from the description in (Rabiner 1989).

The particular use of function-word and suffix sequences has several advantages, the foremost being that no large-scale lexical resources or annotated corpora are required; raw (accented) text is used for training. It is most viable in morphologically rich languages, and may be extended naturally to a full part-of-speech tagger through EM iteration.

The approach exhibits several weaknesses, however. The first is that there are many suffixes in Romance languages, yielding very large matrices and sparse data. It would clearly be desirable to recognize that the suffixes *-aba* and *-ía* both represent the same 1st/3rd person singular imperfect tense (just for different conjugation paradigms) and are functionally equivalent. This and further clustering can be accomplished either manually or by empirical induction. However, a greater problem is the noise introduced when several parts of speech have the same suffix. For example, some nouns may also end in *-aba* and *-ía*, although these are primarily imperfect tense markers. This noise may be tolerable given the relatively low entropy of $p(part\_of\_speech|suffix)$ in Romance languages, but it is apparent that improvement could be achieved using existing dictionary resources to distinguish such cases.

This basic suffix model was implemented with all words assigned automatically to the longest match in a list of 54 suffixes and 40 common function words, with the residual labeled with one of 6 simple classes including punctuation and number. Performance is presented in Table 2 (SUFFIX).

Another variant of Method 2 that was tested here is to use additional dictionary resources in the spirit of (Merialdo, 1990), specifically with the Collins Spanish-English Dictionary and the Liberman-Tzoukermann morphological transducer (1990) used for extrapolation to inflected forms. In this study, the tags are traditional parts of speech (e.g. ADV, ADJ, SPRON, PASTPART), plus individual tags for important function words (e.g. QUE, ...). Suffixes involved in accent ambiguities (-ARA, -ARÁ, -ARAN, -ARÁN, etc.) are given their own tag to allow for specialized context modelling for each of these cases.

For the part-of-speech tags, it is assumed that a word may be tagged with each entry listed in the dictionary for that word with equal probability, with the residual receiving a small epsilon probability. For the suffix tags, the true probability distributions can be extracted from the training corpus. Thus the relative probability of the de-accented *anuncio* being a past-tense verb (*anunció*) or noun (*anuncio*) is directly measured and exploited in classification.

Performance of this specific approach is given in Table 2 (P.O.S.), broken down by general ambiguity type[3]. These results are based on a tagset of 61 parts of speech. Although this variant of Method 2 makes use of dictionary knowledge not available in the suffixes themselves, it uses a smaller tagset (including fewer function words) and makes fewer lexical distinctions, which may explain why it is sometimes outperformed by the suffix-only method.

TABLE 2. N-Gram tagger accuracy

Pattern 1	Pattern 2	SUFFIX	P.O.S.	BaseL	N
anuncio	anunció	97.4%	95.8%	57%	9459
registro	registró	97.0%	97.0%	60%	2596
marco	marcó	97.8%	97.5%	52%	2069
completo	completó	92.6%	85.2%	54%	1701
retiro	retiró	97.0%	97.3%	56%	3713
duro	duró	90.3%	93.7%	52%	1466
paso	pasó	88.0%	93.9%	50%	6383
regalo	regaló	89.5%	89.5%	56%	280
terminara	terminará	60.0%	65.7%	59%	218
llegara	llegará	68.9%	67.6%	64%	860
esta	está	88.7%	85.8%	61%	14140
mi	mí	90.0%	94.1%	82%	1221
secretaria	secretaría	52.3%	52.3%	52%	1065

There is considerable opportunity for further improvement in this approach. A hand-tagged corpus could be used for better initial probability

[3]The words used in this comparative study are a random selection from the most problematic cases of each ambiguity type – those exhibiting the largest absolute number of the non-majority accent patterns. Collectively they are representative of the most common potential sources of error. The training and test sets were independent in all cases, and the examples were extracted from the Spanish AP Newswire (1991-1993, 49 million words). These same words were also used to test all the other methods in this comparative study.

estimation, and the EM algorithm could be used to refine B-Matrix probabilities iteratively (Merialdo, 1990). However, the goal of this study was not to produce a full part-of-speech tagger, but to improve ambiguity resolution in accent restoration. A cost-benefit analysis could help determine whether additional resources are worth devoting to this approach.

This method has several fundamental limitations for the task of accent restoration. First, it is not adequately lexicalized. For example, for the *-ara/-ará* subjunctive/future distinction, the presence of temporal words (days of the week, months, events, etc.) are highly significant, and for other tense distinctions specific lexical associations are important. One could add additional word classes, but there are many more useful distinctions than can be adequately accommodated given the algorithm's time and space complexity bounds. More intractably, however, many of the necessary tense distinctions are sensitive to mid-to-long distance word associations (such as the temporal indicators) that simply cannot be captured with an n-gram model, for any reasonable size of n. And finally, the approach does not address the cases that arise when a token has multiple accented forms with the same part of speech. It would appear that further progress can best be made by developing more lexicalized and longer-distance models of context.

2.3. METHOD 3: BAYESIAN CLASSIFIER

Bayesian classifiers are particularly well suited for handling highly lexicalized and longer-distance models of context, two of the central weaknesses of the previous approach. They have been employed successfully in word-sense disambiguation (Gale, Church and Yarowsky, 1992), authorship identification (Mosteller and Wallace, 1964) and person-place classification of proper nouns (Gale, Church and Yarowsky, 1994).

The basic technique employed is to treat a window of words surrounding each ambiguous word as a document, and ask if there are any measurable differences in the distribution of words found in the contexts surrounding one of its accent patterns relative to the other. For example, when distinguishing the accent patterns *terminara* (subjunctive) from *terminará* (future), one would tend to find that the token *domingo* (Sunday) occurs much more frequently in the context of the latter than the former, while certain subjunctive marking phrases occur in an inverse distribution. Considering only one of these words in context, we can estimate the probability that the context belongs to one accent pattern relative to another by the likelihood ratio:

$$\frac{p(token\_in\_context|accent\_pattern_1)}{p(token\_in\_context|accent\_pattern_2)}$$

Making the simplifying assumption that all tokens seen in the context of an ambiguous word provide *independent* evidence for classifying the accent pattern, we can combine the ratios in a product to yield an overall likelihood ratio that the ambiguous token has one accent pattern relative to another:

$$\prod_{token \text{ in } context} \frac{p(token_i|accent\_pattern_1)}{p(token_i|accent\_pattern_2)}$$

A primary variable here is the width of context considered. Two experiments were conducted, one examining a fairly wide context (± 20 words) and one examining a more localized context ($\pm 2-4$ words). The larger is similar to the width often employed in sense disambiguation, and is useful for modelling "semantic" or "topic" differences, while the smaller window is better suited for modelling more "syntactic" distinctions.

Table 3 provides an outline of the performance of Method 3 (Bayesian Classifiers), using context window sizes of ± 2, ± 4 and ± 20 words.

TABLE 3. Bayesian classifier accuracy

Pattern 1	Pattern 2	±2	±4	±20	BaseL
anuncio	anunció	85.5	88.4	74.7	57 %
registro	registró	87.1	81.8	77.0	60 %
marco	marcó	94.4	93.0	93.5	52 %
completo	completó	90.6	89.2	88.6	54 %
retiro	retiró	88.1	88.6	79.3	56 %
duro	duró	93.5	93.4	82.1	52 %
paso	pasó	88.2	86.5	76.4	50 %
regalo	regaló	84.7	80.4	75.9	56 %
terminara	terminará	79.2	83.5	82.8	59 %
llegara	llegará	65.6	72.2	62.9	64 %
esta	está	88.2	87.8	81.3	61 %
mi	mí	80.4	79.9	76.7	76 %
secretaria	secretaría	78.1	75.0	75.6	52 %

The Bayesian classifier has the advantage of not requiring special lexical resources or annotated corpora. It supports a highly lexicalized feature set and may capture long-distance dependencies. It can distinguish ambiguities within the same part of speech.

However, the major disadvantage of the "bag of words" Bayesian classifier approach is that it is difficult to model the occurrence of words in specific positions. Given the assumption of independence, it is also quite difficult to model *sequences* of nearby words, when the joint appearance of two or more words differ in their distribution from that expected from the product of their independent likelihood ratios. This independence assumption also makes the technique poorly suited for combining multiple non-independent sources of evidence, such as parts-of-speech, lemmas, word classes and individual inflected words all in the same context.

2.4. METHOD 4: DECISION LISTS

The limitations observed above are precisely what has motivated the development of Method 4, a hybrid approach using decision lists, combining the strengths of both Bayesian classifiers and N-gram taggers. This approach was derived from the formal model of decision lists

presented in (Rivest, 1987). However, feature conjuncts have been restricted to a much narrower complexity than allowed in the original model – namely to word and class trigrams. Early results presented in (Sproat, Hirschberg and Yarowsky, 1992) achieved 97% mean accuracy on the problem of homograph resolution in text-to-speech synthesis[4]. The current approach was proposed in (Yarowsky, 1994) and is described more fully there. Below is an outline of the algorithm:

Steps 1 & 2: Measure Accent Pattern Distributions and Collect Training Contexts
The algorithm begins by identifying the accent pattern ambiguities for a language. An accent distribution table is computed as described in Method 1 (Baseline). For each case of accent ambiguity identified, collect $\pm k$ words of context around all occurrences in the training corpus, label the concordance line with the observed accent pattern, and then strip the accents from the data. This will yield a training set such as in Table 4.

Step 3: Measure Collocational Distributions
The driving force behind this disambiguation algorithm is the uneven distribution of collocations[5] with respect to the ambiguous token being classified. The presence of certain collocations will indicate one accent pattern, while different collocations will tend to indicate another. The goal of this stage of the algorithm is to measure a large number of collocational distributions and select those which are most useful in identifying the accent pattern of the ambiguous word.

[4]For the data set of 13 homographs used in this study, baseline correctness was 67%.
[5]The term *collocation* is used here in its broad sense, meaning words appearing adjacent to or near each other (literally, in the same location), and should not imply only idiomatic or non-compositional associations.

TABLE 4. Sample training data

Pattern	Context		
(1) côté	du laisser de	*cote*	faute de temps
(1) côté	appeler l' autre	*cote*	de l' atlantique
(1) côté	passe de notre	*cote*	de la frontiere
(2) côte	vivre sur notre	*cote*	ouest toujours verte
(2) côte	creer sur la	*cote*	du labrador des
(2) côte	travaillaient cote a	*cote*	, ils avaient

The following are the initial types of collocations considered:

- Word immediately to the left (-1 W)
- Word found in $\pm k$ word window[6] (±k W)
- Pair of words at offsets -2 and -1
- Pair of words at offsets -1 and +1
- Pair of words at offsets +1 and +2

For the two major accent patterns of the French noun *cote*, Table 5 shows a small sample of these distributions for several types of collocations.

By themselves, such simple word associations have considerable discriminating power, and can successfully model gender constraints, etc. without these constraints being explicitly represented (or known). However, if additional resources such as a morphological analyzer are available, similar collocational patterns for linguistic features such as morphological root may be measured. This often yields more succinct and generalizable discriminators than achieved from a list of the observed inflected forms. The Tzoukermann/Liberman (1990) Spanish morphological analyzer was used here for this purpose. Similarly, distributional patterns for part-of-speech bigrams and trigrams were computed, using a relatively coarse level of analysis (such as NOUN, ADJECTIVE, SUBJECT-PRONOUN, ARTICLE, etc.) comparable to that used in Method 2. However, since the information was extracted from a dictionary and not from a part-of-speech-tagged corpus, no relative frequency distribution was available for words with multiple parts-of-speech. Such words were given a part-of-speech tag consisting of the union of the possibilities (eg ADJECTIVE-NOUN), as in Kupiec (1989). Thus sequences of pure part-of-speech tags were highly reliable, while the potential sources of noise were isolated and modeled separately. In addition, collocational

[6]The optimal value of k is sensitive to the type of ambiguity. Semantic or topic-based ambiguities warrant a larger window ($k \approx 20-50$), while more local syntactic ambiguities warrant a smaller window ($k \approx 3$ or 4)

TABLE 5. A sample of collocational patterns used in classification

Position	Collocation	**côte**	**côté**
-1 w	du *cote*	0	536
	la *cote*	766	1
	un *cote*	0	216
	notre *cote*	10	70
+1 w	*cote* ouest	288	1
	cote est	174	3
	cote du	55	156
+1w,+2w	*cote* du gouvernement	0	62
-2w,-1w	cote a *cote*	23	0
±k w, $k = 20$	poisson (within ±20 words)	20	0
±k w, $k = 20$	ports (within ±20 words)	22	0
±k w, $k = 20$	opposition (within ±20 words)	0	39

statistics were measured for several common word classes, such as WEEKDAY (={ *domingo*, *lunes*, *martes*, ... }) or MONTH, primarily focusing on time words because so many accent ambiguities involve tense distinctions.

To build a full part of speech tagger for Spanish would be quite costly (and require special tagged corpora). The current approach uses only the information available in dictionaries, exploiting only that which is useful for the accent restoration task. Were dictionaries not available, a productive approximation could have been made using the associational distributions of suffixes (such as *-aba*, *-aste*, *-amos*) which are often satisfactory indicators of part of speech in morphologically rich languages such as Spanish.

For the French experiments, no additional linguistic knowledge or lexical resources were used. The decision lists were trained solely on raw word associations without additional patterns based on part of speech, morphological analysis or word class. Hence the reported performance is representative of what may be achieved with a rapid, inexpensive implementation based strictly on the distributional properties of raw text.

The use of the word-class and part-of-speech data is illustrated in Table 6, with the example of distinguishing *terminara/terminará* (a subjunctive/future tense ambiguity).

Step 4: Sort by Log-Likelihood into Decision Lists

For each individual collocation, the following log-likelihood ratio was computed:

TABLE 6. Use of word-class and part-of-speech patterns

Position	Collocation	**terminara**	**terminará**
-2P,-1P	PREPOSITION QUE *terminara*	31	0
-2W,-1W	de que *terminara*	15	0
-2W,-1W	para que *terminara*	14	0
-2P,-1P	NOUN QUE *terminara*	0	13
-2W,-1W	carrera que *terminara*	0	3
-2W,-1W	reunion que *terminara*	0	2
-2W,-1W	acuerdo que *terminara*	0	2
-1W	que *terminara*	42	37
±k C, $k = 20$	WEEKDAY (within ±20 words)	0	23
±k W, $k = 20$	domingo (within ±20 words)	0	10
±k W, $k = 20$	viernes (within ±20 words)	0	4

$$Abs(Log(\frac{p(Accent\_Pattern_1|Collocation_i)}{p(Accent\_Pattern_2|Collocation_i)}))$$

The collocations most strongly indicative of a particular pattern will have the largest log-likelihood. Sorting by this value will list the strongest and most reliable evidence first[7]

Evidence sorted in the above manner will yield a decision list like the following, highly abbreviated example[8] in Table 7.

The resulting decision list is used to classify new examples by identifying the highest line in the list that matches the given context and returning the indicated classification. The algorithm differs markedly here from the Bayesian classifier and N-gram tagger in that it does *not* combine the scores for each member of the list found in the target context to be tagged, but

[7]Problems arise when an observed count is 0. Clearly the probability of seeing *côté* in the context of *poisson* is not 0, even though no such collocation was observed in the training data. Finding a more accurate probability estimate depends on several factors, including the size of the training sample, nature of the collocation (adjacent bigrams or wider context), our prior expectation about the similarity of contexts, and the amount of noise in the training data. Several smoothing methods have been explored here, including those discussed in (Gale *et al.*, 1992). In one technique, all observed distributions with the same 0-denominator raw frequency ratio (such as 2/0) are taken collectively, the average agreement rate of these distributions with additional held-out training data is measured, and from this a more realistic estimate of the likelihood ratio (e.g. 1.8/0.2) is computed. See Yarowsky (1996) for further description of this approach.

[8]Entries marked with † are pruned in Step 5, below.

TABLE 7. A sample, highly abbreviated decision list

LogL	Evidence	Classification
8.28	PREPOSITION QUE *terminara*	⇒ terminara
†7.24	de que *terminara*	⇒ terminara
†7.14	para que *terminara*	⇒ terminara
6.87	y *terminara*	⇒ terminará
6.64	WEEKDAY (within ±20 words)	⇒ terminará
5.82	NOUN QUE *terminara*	⇒ terminará
†5.45	domingo (within ±20 words)	⇒ terminará

rather uses only the single best piece of evidence available. See Step 7 for a discussion of this process.

Step 5: Interpolation and Pruning

A potentially useful optional procedure is the interpolation of log-likelihood ratios between those computed from the full data set (the *global* probabilities) and those computed from the residual training data left at a given point in the decision list when all higher-ranked patterns failed to match (i.e. the *residual* probabilities). The residual probabilities are more relevant, but since the size of the residual training data shrinks at each level in the list, they are often much more poorly estimated (and in many cases there may be no relevant data left in the residual on which to compute the distribution of accent patterns for a given collocation). In contrast, the global probabilities are better estimated but less relevant. A reasonable compromise is to interpolate between the two where the interpolated estimate is $\beta_i \times global + (1 - \beta_i) \times residual$. When the residual probabilities are based on a large training set and are well estimated, β_i is small (the residual will dominate). In cases where the relevant residual is small or non-existent, β_i will be large and the smoothed probabilities will rely primarily on the better estimated global values. If all $\beta_i = 0$ (exclusive use of the residual), the result is a degenerate (strictly right-branching) decision tree with severe sparse data problems. Alternately, if one assumes that likelihood ratios for a given collocation are functionally equivalent at each line of a decision list, then one could exclusively use the global (all $\beta_i = 1$). This is clearly the easiest and fastest approach, as probability distributions do not need to be recomputed as the list is constructed. Which approach is best? Using only the global proabilities does surprisingly well, and the results cited here are based on this readily replicable procedure. The reason is grounded in the strong tendency of a word to exhibit only one sense or accent pattern per collocation (discussed in Step 7 and (Yarowsky, 1993)). Most classifications are based on a x vs. 0 distribution, and while the magnitude of the log-likelihood ratios may decrease in the residual, they rarely change sign. There are cases where this does happen and it appears that some interpolation helps, but for *this* problem the relatively small difference in performance does not seem to justify the greatly increased computational cost.

Two kinds of optional pruning can also increase the efficiency of the decision lists. The first handles the problem of "redundancy by subsumption," which is visible in the decision lists above (in WEEKDAY and *domingo*). When lemmas and word-classes precede their member words in the list, the latter will be ignored and can be pruned. If a bigram is unambiguous, probability distributions for dependent trigrams will not even be generated, as they provide no additional information.

The second, pruning in a cross-validation phase, compensates for the minimal observed over-modeling of the data. Once a decision list is built it is applied to its own training set plus some held-out cross-validation data (*not* the test data). Lines in the list which contribute to more incorrect classifications than correct ones are removed. This also indirectly handles problems that may result from the omission of the interpolation step. If space is at a premium, lines which are never used in the cross-validation step may also be pruned. However, useful information is lost here, and words pruned in this way may have contributed to the classification of testing examples. A 3% drop in performance is observed, but an over 90% reduction in space is realized. The optimum pruning strategy is subject to cost-benefit analysis. In the results reported below, all pruning except this final space-saving step was utilized.

Step 6: Train Decision Lists for General Classes of Ambiguity

For many similar types of ambiguities, such as the Spanish subjunctive/future distinction between *-ara* and *ará*, the decision lists for individual cases will be quite similar and use the same basic evidence for the classification (such as presence of nearby time adverbials). It is useful to build a general decision list for all *-ara/ará* ambiguities. This also tends to improve performance on words for which there is inadequate training data to build a full individual decision lists. The process for building this general class disambiguator is basically identical to that described in Steps 2-5 above, except that in Step 2, training contexts are pooled for all individual instances of the class (such as all *-ara/-ará* ambiguities). It is important to give each individual *-ara* word roughly equal representation in the training set, however, lest the list model the idiosyncrasies of the most frequent class members, rather than identify the shared common features representative of the full class.

In Spanish, decision lists are trained for the general ambiguity classes including *-o/-ó*, *-e/-é*, *-ara/-ará*, and *-aran/-arán*. For each ambiguous word belonginging to one of these classes, the accuracy of the word-specific decision list is compared with the class-based list. If the class's list performs adequately it is used. Words with idiosyncrasies that are not modeled well by the class's list retain their own word-specific decision list .

Step 7: Using the Decision Lists

Once these decision lists have been created, they may be used in real time to determine the accent pattern for ambiguous words in new contexts.

At run time, each word encountered in a text is looked up in a table. If the accent pattern is unambiguous, as determined in Step 1, the correct pattern is printed. Ambiguous words have a table of the possible accent patterns and a pointer to a decision list , either for that specific word or its ambiguity class (as determined in Step 6). This given list is searched for the highest ranking match in the word's context, and a classification number is returned, indicating the most likely of the word's accent patterns given the context[9].

From a statistical perspective, the evidence at the top of this list will most reliably disambiguate the target word. Given a word in a new context to be assigned an accent pattern, if we may only base the classification on a single line in the decision list , it should clearly be the highest ranking pattern that is present in the target context.

The question, however, is what to do with the less-reliable evidence that may also be present in the target context. The common tradition is to combine the available evidence in a weighted sum or product. This is done by Bayesian classifiers, neural nets, IR-based classifiers and N-gram part-of-speech taggers. The system reported here is unusual in that it does no such combination. *Only* the single most reliable piece of evidence matched in the target context is used. For example, in a context of *cote* containing *poisson*, *ports* and *atlantique*, if the adjacent feminine article *la cote* (the coast) is present, only this best evidence is used and the supporting semantic information ignored. Note that if the masculine article *le cote* (the side) were present in a similar maritime context, the most reliable evidence (gender agreement) would override the semantic clues which would otherwise dominate if all evidence was combined. If no gender agreement constraint were present in that context, the first matching semantic evidence would be used.

[9] If all entries in a decision list fail to match in a particular new context, a final entry called DEFAULT is used; it indicates the most likely accent pattern in cases where nothing matches.

There are several motivations for this approach. The first is that combining all available evidence rarely produces a different classification than just using the single most reliable evidence, and when these differ it is as likely to hurt as to help. In a study comparing results for 20 words in a binary homograph disambiguation task, based strictly on words in local (± 4 word) context, the following differences were observed between an algorithm taking the single best evidence, and an otherwise identical algorithm combining all available matching evidence:

TABLE 8. Combining vs. not combining probabilities

Agree -	Both classifications correct	92%
	Both classifications incorrect	6%
Disagree -	Single best evidence correct	1.3%
	Combined evidence correct	0.7%
Total -		100%

Of course that this behavior does not hold for all classification tasks, but *does* seem to be characteristic of lexically-based word classifications. This may be explained by the empirical observation that in most cases, and with high probability, words exhibit only one *sense* in a given collocation (Yarowsky, 1993). Thus for this type of ambiguity resolution, there is no apparent detriment, and some apparent performance gain, from using only the single most reliable evidence in a classification. There are other advantages as well, including run-time efficiency and ease of parallelization. However, the greatest gain comes from the ability to incorporate multiple, non-independent information types in the decision procedure. As noted above, a given word in context may match several times in the decision list , once for its part of speech, lemma, inflected form, trigrams, and possibly word-class as well. By only using one of these matches, the gross exaggeration of probability from combining all of these non-independent log-likelihoods is avoided. While these dependencies may be modeled and corrected for in Bayesian formalisms, it is difficult and costly to do so. Using only one log-likelihood ratio without combination frees the algorithm to include a wide spectrum of highly non-independent information without additional algorithmic complexity or performance loss.

Evaluation
Table 9 below gives a breakdown of performance on the comparative test set[10]. All of these evaluations were conducted with 5-fold cross-validation, using independent training and testing data.

TABLE 9. Decision list accuracy on specific accent-pattern ambiguities

Spanish:				
Pattern 1	Pattern 2	Agreement	BaseL	N
anuncio	anunció	98.4%	57%	9459
registro	registró	98.4%	60%	2596
marco	marcó	98.2%	52%	2069
completo	completó	98.1%	54%	1701
retiro	retiró	97.5%	56%	3713
duro	duró	96.8%	52%	1466
paso	pasó	96.4%	50%	6383
regalo	regaló	90.7%	56%	280
terminara	terminará	82.9%	59%	218
llegara	llegará	78.4%	64%	860
esta	está	97.1%	61%	14140
mi	mí	93.7%	82%	1221
secretaria	secretaría	84.5%	52%	1065
French:				
cessé	cesse	97.7%	53%	1262
décidé	décide	96.5%	64%	3667
laisse	laissé	95.5%	50%	2624
commence	commencé	95.2%	54%	2105
côté	côte	98.1%	69%	3893
traité	traite	95.6%	71%	2865

3. Comparative Evaluation

A comparative analysis of system performance on the major ambiguity types found in Spanish is provided in Table 10. The numbers are an average

[10]The French results are presented for reference only. Although all three algorithms have been applied to French data, space has limited the comparative evaluation to Spanish.

of the results for the test set of words presented in the preceding tables[11]. The common set of test cases helps highlight differences in performance between the four algorithms.

TABLE 10. Comparison of accuracy on Spanish

Type	# 1 Baseline	# 2a N-gram (suffix)	# 2b N-gram (P.O.S.)	# 3 Bayesian Classifier	# 4 Decision List
-o/-ó	54.6	93.7	93.8	89.4	96.8
-ara/-ará	61.5	64.4	66.6	77.1	80.6
Function Word	81.8	89.3	89.9	84.3	95.4
Same POS	52.0	52.3	52.3	78.1	84.5

These results confirms several earlier hypotheses. First, the N-gram tagger and decision list are the best discriminators for *-o/-ó* and function word ambiguities, which involve primarily local, syntactic distinctions. Bayesian classifiers are less well suited for this task. In contrast, *-ara/-ará* ambiguities (of tense and mood), which involve longer range semantic dependencies, are best handled by decision lists followed by Bayesian classifiers. N-gram taggers perform very poorly on this task, as the distinguishing evidence is often beyond the immediate three word window. For ambiguities involving two words of the same part of speech, Bayesian classifiers and decision lists also perform best, while the part-of-speech-based N-gram tagger is not able to handle this case at all. Thus while the N-gram tagger and Bayesian classifier perform well on complementary subsets of the problem, the decision list algorithm performs well on both. It offers generality without apparent loss of precision.

Further analysis of these differences is presented below. However, before continuing with further comparison, it is important to note that all precision values in these experiments are based on *agreement* rates with the accent patterns in the test set, which themselves may be erroneous. Because accents have only been stripped artificially for testing purposes, and the "correct" patterns exist on-line in the original corpus, it is entirely objective and automatic to test performance, unlike the evaluation of word-sense disambiguation and part-of-speech tagging, where at some point human judgements are required. Regrettably, however, there are errors in

[11]In the case of the Bayesian Classifier, the performance from the optimal context width is used. Note that this context range was typically ± 2, the same as used by a trigram model.

the original corpus, which can be quite substantial depending on the type of accent. For example, accents over the i (í) are frequently omitted, and in a sample test 3.7% of the appropriate accents were missing. Thus the previous results must be interpreted as agreement rates with the corpus accent pattern; the true percent correct may be several percentage points higher. The relatively low agreement rate on words with accented i's (í) in Tables 2, 3, 9 are a result of these corpus errors. To study this discrepancy further, a human judge fluent in Spanish determined whether the corpus or decision list algorithm was correct in instances of two common ambiguities. For the ambiguity case of *mi/mí*, the corpus was incorrect in 46% of the disputed tokens. For the ambiguity *anuncio/anunció*, the corpus was incorrect in 56% of the disputed tokens. Although these appear to be extreme cases, they indicate limits to the precision of automatic evaluation.

At some point it would be interesting to pursue a more comprehensive evaluation of the accuracy of the corpus relative to the algorithm, and get a more precise estimate of algorithm "correctness" rather than just agreement with the corpus. However, this would require considerable effort, including multiple human judges and other mechanisms to reduce bias. Since automatic, objective scoring was one of the primary motivations for using the corpus accentings as the evaluation gold-standard, I would hope to find a more reliable corpus of test material instead. Nevertheless, initial results indicate that in some cases, the decision list system's precision may rival that of the AP Newswire's Spanish writers, translators, and copy-editors.

4. Discussion and Conclusions

The decision list algorithm presented here combines the strengths of both N-gram taggers and Bayesian classifiers, and outperforms both. Like the N-gram tagger, it utilizes trigram probabilities to model local syntactic constraints, and like the Bayesian classifier it successfully models long range lexical associations of a more semantic nature. By incorporating multiple types of evidence, the decision list not only exhibits generality and the ability to perform well on very different types of disambiguation problems, it also can exploit the additional available information to outperform the two competing algorithms on the tasks for which they are specialized.

It is often difficult to know in advance what information will be most useful for a particular discrimination task. Decision lists *consider* an extremely broad set of evidence in the training phase, but only utilize that which is most effective as a discriminating agent for the given task. While one could incorporate multiple sources of evidence in a Bayesian classifier as well, the key advantage of this decision list algorithm is that it allows the use of multiple, highly non-independent evidence types (such as root form, inflected form, part of speech, thesaurus category or application-specific clusters) and does so in a way that avoids the complex modelling of statistical dependencies. This allows the decision lists to find the level of representation that best matches the observed probability distributions. It is a kitchen-sink approach of the best kind – throw in many types of potentially relevant features and watch what floats to the top.

While there are certainly other ways to combine such evidence, the decision list approach has many advantages. The foremost is simplicity – it is extremely straightforward to use any new feature for which a probability distribution can be calculated. The algorithm, especially in its most basic form, is very easy to describe and implement. Other advantages are its perspicuity: the decision list is organized like a recipe, with the most useful evidence first and in highly readable form. It is much more comprehendable than an N-gram matrix or one of the impenetrable black boxes produced by many other machine learning algorithms. The generated decision procedure is easy to edit by hand, changing or adding patterns to the list. The algorithm is also readily applied to new domains: it was originally developed for homograph disambiguation in text-to-speech synthesis (Sproat *et al.*, 1992), and was applied to the current problem without modification. The flexibility and generality of the algorithm and its potential feature set makes it readily applicable to other problems of recovering lost information from text corpora such as the tasks of capitalization restoration and recovering vowels in Hebrew text.

Overall, the decision list algorithm demonstrates considerable hybrid vigor, combining the strengths of N-gram taggers and Bayesian classifiers in a highly effective, general purpose decision procedure for lexical ambiguity resolution.

References

Church, K.W. (1988) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, in *Proceedings of the Second Conference on Applied Natural Language Processing, ACL*, 136-143.

Gale, W., K. Church, and D. Yarowsky (1992), A Method for Disambiguating Word Senses in a Large Corpus, *Computers and the Humanities*, 26, 415-439.

Gale, W., K. Church, and D. Yarowsky (1994), Discrimination decisions for 100,000-dimensional spaces. In A. Zampoli, N. Calzolari and M. Palmer (eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker*, Kluwer Academic Publishers, pp. 429–450.

Kupiec, Julian (1989) Probabilistic Models of Short and Long Distance Word Dependencies in Running Text, in *Proceedings, DARPA Speech and Natural Language Workshop*, Philadelphia, February, pp. 290-295.

Leacock, Claudia, Geoffrey Towell and Ellen Voorhees (1993) Corpus-Based Statistical Sense Resolution, in *Proceedings, ARPA Human Language Technology Workshop.*

Merialdo, B. (1990) 'Tagging Text with a Probabilistic Model, in *Proceedings of the IBM*

Natural Language ITL, Paris, France, pp. 161-172.

Mosteller, Frederick, and David Wallace (1964) *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Massachusetts.

Paul, D. B. (1990) Speech Recognition Using Hidden Markov Models, in *The Lincoln Laboratory Journal*, 3.

Rabiner, L. R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, in *Proceedings of the IEEE*, 77, 257-285.

Rivest, R. L. (1987) Learning Decision Lists, in *Machine Learning*, 2, 229-246.

Sproat, R., J. Hirschberg and D. Yarowsky (1992) A Corpus-based Synthesizer, in *Proceedings, International Conference on Spoken Language Processing*, Banff, Alberta. October 1992.

Tzoukermann, Evelyne and Mark Liberman (1990) A Finite-state Morphological Processor for Spanish, in *Proceedings, COLING-90*, Helsinki.

Yarowsky, David (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, in *Proceedings, COLING-92*, Nantes, France.

Yarowsky, David (1993) One Sense Per Collocation, in *Proceedings, ARPA Human Language Technology Workshop*, Princeton.

Yarowsky, David (1994) Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM.

Yarowsky, David (1995) Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, pp. 189–196.

Yarowsky, David (1996) Homograph Disambiguation in Speech Synthesis, in J. Hirschberg, R. Sproat and J. van Santen (eds.), *Progress in Speech Synthesis*, Springer-Verlag.

INDEX