Pronunciation Disambiguation in Turkish

M. Oğuzhan Külekci[1,2] and Kemal Oflazer[1]

[1] Faculty of Engineering and Natural Sciences, Sabancı University
Tuzla, Istanbul, Turkey 34956
[2] TÜBİTAK-UEKAE, Gebze, Kocaeli, Turkey 41470
kulekci@su.sabanciuniv.edu, oflazer@sabanciuniv.edu

Abstract. In text-to-speech systems and in developing transcriptions for acoustic speech data, one is faced with the problem of disambiguating the pronunciation of a token in the context it is used, so that the correct pronunciation can be produced or the transcription uses the correct set of phonemes. In this paper we investigate the problem of pronunciation disambiguation in Turkish as a natural language processing problem and present preliminary results using a morphological disambiguation technique based on the notion of *distinguishing tag sets*.

1 Introduction

Words typically have different pronunciations depending on their syntactic, and semantic properties in context. In Turkish, differences in pronunciation stem from differences in the phonemes used, the length of the vowel and the location of the primary stress [1]. The selection of the correct pronunciation requires a disambiguation process that needs to look at local morphosyntactic and semantic information to determine the correct pronunciation among alternatives. Disambiguating morphology serves a good starting basis for disambiguating of pronunciations, although it by itself, does not disambiguate all ambiguous cases of pronunciation. For example, determining the correct morphological analysis of the word *okuma* in Turkish, distinguishes between the possible pronunciations of this word in the sentences *Okuma kitabı belirlendi.* 'Reading book has been determined.' and *Saçma sapan şeyleri okuma.* 'Don't read those silly things.' In the former, *okuma* is an infinitive form derived from verb *okumak* (to read) and corresponds to phonetic representation /o-ku-"ma/ in SAMPA representation.[1,2] In the latter case the same word functions as an imperative form of the same verb, and pronunciation is represented with /o-"ku-ma/ where the primary

[1] SAMPA(Speech Assessment Methods Pronunciation Alphabet) is an international machine-readable pronunciation alphabet. For further information, please refer to `www.phon.ucl.ac.uk/home/sampa`. See `http://www.phon.ucl.ac.uk/home/sampa/turkish.htm` for the set of Turkish SAMPA phoneme representations. We use the SAMPA notation to represent pronunciations in the text, where necessary.

[2] "Indicates the stressed syllable, and - indicates a syllable boundary.

P. Yolum et al.(Eds.): ISCIS 2005, LNCS 3733, pp. 636–645, 2005.

stress is on the second syllable. A text-to-speech system would have to take this into account for proper prosody.

Morphological disambiguation has employed stochastic, knowledge-based, and hybrid methods ([2,3,4,5,6]). Morphological disambiguation for Turkish has experimented with various alternatives: Oflazer and Kuruöz [7] have used a symbolic approach using handcrafted rules. Oflazer and Tür [3] have proposed a similar scheme with rules learned automatically. Oflazer and Tür also have used voting constraints [8]. Recently, statistical morphological disambiguation of Turkish has been studied by Tür *et al.* [6]. None of these however have been applied to the problem of disambiguating pronunciation. We should also mention in passing that full morphological disambiguation is an *overkill* for disambiguating pronunciations, since in general words have less pronunciation ambiguity than morphological ambiguity.

2 Pronunciation Ambiguities in Turkish

Turkish orthography, uses 29 letters to encode its orthography but phonologically there are 34 phonemes: the 8 vowels /i, y, e, 2, a, o, 1, u/ which correspond to *i*, *ü*, *e*, *ö*, *a*, *o*, *ı*, and *u* in orthography and the 26 consonants: /p, t, tS, k, c, b, d, dZ, g, gj, f, s, S, v, w, z, Z, m, n, N, l, 5, r, j, h, G/. Orthography uses only 21 letters for consonants: /g/ and its palatal counterpart /gj/ are written as *g*, while /k/ and and its palatal counterpart /c/ are written as *k*, /5/ and its palatal counterpart /l/ are written as *l*, /v, w/ are written as *v*, and /n/ and its nasal counterpart /N/ are written as *n*. Palatalized segments (/gj, c, l/) contrast with their nonpalatalized counterparts only in the vicinity of back vowels (thus *sol* is pronounced /so5/ when used to mean ‘left’ vs. /sol/ when used to mean ‘note in scale’). In the neighborhood of front vowels, palatality is predictable (*lig* /ligj/ ‘league’).[3] /G/, written as *ğ*, represents the velar fricative or glide corresponding to the historical voiced velar fricative that was lost in Standard Turkish. When it is syllable-final, some speakers pronounce it as a glide and others just lengthen the preceding vowel. We treat it as a consonant for the purposes of this work and explicitly represent it. This inventory does not include long vowels – such phonemes are indicated with a vowel length symbol.[4]

Statistics in Table 1 from Oflazer and Inkelas [1] over a 1 million corpus indicate that approximately 90% of the words in running Turkish text have single, 9% have two and only 1% has more than two distinct pronunciations. Thus, a Turkish TTS synthesizer would have to resolve ambiguities in 10% of the words in an input text.

[3] In conservative spellings of some words, contrastive velar or lateral palatality is indicated with a circumflex on the adjacent vowel, though this convention actually ambiguous and because circumflexes are also used in some words, equally sporadically, to indicate vowel length.

[4] It is certainly possible to come up with a finer set of phonemes especially for text-to-speech purposes, so that the effects of palatal consonants, etc., can be distributed to the neighboring vowels.

Table 1. Statistics about the pronunciation ambiguity observed in Turkish

Average Morphological Parse-Pronunciation Pairs / Token	1.86
Average Distinct Morphological Parses / Token	1.84
Average Distinct Pronunciations / Token	1.11
Average Distinct Pronunciations (ignoring stress) / Token	1.02

There are three types of pronunciation ambiguities in Turkish arising from (i) the phonemes used, (ii) the position of the primary stress, and (iii) differences in vowel length. The numbers in Table 1 indicate that the main source of ambiguity to be resolved in Turkish pronunciation is the position of the primary stress and if we ignore the position of the stress only 2% of the tokens has ambiguities such as differences in vowel length and consonant palatality in the root portions of the words.

These differences in pronunciations manifest themselves in various combinations, and different techniques have to be applied to resolve the resulting ambiguities:

1. The root words are homographs but have different parts-of-speech: *ama*(/"a-ma/, `ama+Conj`, 'but') vs. *ama* (/a:-"ma:/, `ama+Adj`, 'blind'). Morphological disambiguation would be able to resolve such ambiguities.
2. The root words are homographs *and* have the same part-of-speech; and further they inflect in exactly the same way: *kar* (/"kar/ 'snow') vs. *kar* (/"car/ 'profit') or *yar* (/"jar/ 'ravine') versus *yar* (/"ja:r/ 'lover'). This is akin to the disambiguation in English of *bass* ('fish') vs. *bass* ('musical instrument'). Morphological disambiguation would not be of much use here and one would have to resort to techniques used in word sense disambiguation.
3. The root words are homographs and have the same part of speech and pronounced the same, but under certain inflections, the root word with a certain sense undergoes further changes: For example for the word *hal* (/"hal/ 'fruit market' or 'state'), with the dative case marker suffix we get *hale* (/ha-"le/ `hal+Noun...+Dat`) with the first sense vs. *hale* (/ha:-"le/ `hal+Noun...+Dat` with the second sense (and an additional reading *hale* (/ha:-"le/ `hale+Noun...+Nom` 'halo')). We need to first disambiguate morphology here. If we predict that the word has nominative case, then we know the pronunciation and we are done. However, if we predict that the word has dative case, we now have to resort to word sense disambiguation to select the appropriate pronunciation depending on the sense of the root *hal* is used.
4. The words are homographs but morphological analysis produces multiple segmentations giving rise to free and bound morphemes with different semantics, morphosyntactic functions and stress marking properties: Here are some interesting examples:
   - *ajanda* (/a-"Zan-da/, `ajanda+Noun...+Nom`, 'agenda') vs. *ajanda* (/a-Zan-"da/, `ajan+Noun...+Loc` 'on the agent'). Here the first parse has a root word with exceptional root stress.
   - *fazla* (/faz-"5a/ `fazla+Adverb` 'much') vs. *fazla* (/"faz-5a/ `faz+Noun... +Ins` 'with the phase'). Here, the instrumental case marking morpheme

(-*la*) is prestressing, but happens to surface as the last two phonemes of the first root word.

- *uyardı* (/u-”jar-d1/ `uy+Verb...+Aor+Past+A3sg` ’s/he/it used to fit’) vs. *uyardı* (/u-jar-”d1/ `uyar+Verb...+Past+A3sg` ’s/he warned’). In the first interpretation, the morpheme marking past tense is prestressing when preceded by the aorist aspect morpheme, but not otherwise.
- *attı* (/”at-t1/ `at+Noun...^ DB+Verb+Past+A3sg` ’it was a horse’) vs. *attı* (/at-”t1/ `at+Verb...+Past+A3sg` ’he threw’). Similar to above, in the first interpretation, the morpheme marking past tense is prestressing when applied to a noun or adjective root (through an implicit verbal derivation.)

Most such cases can be resolved with morphological disambiguation.

5. Proper nouns especially those denoting place names that are homographs with common nouns (inflected or otherwise) usually have non-final stress in the root affecting the stress properties of their inflected versions: e.g., *Ordu* (/”or-du/ ’name of a city’) versus *Ordu* (/or-”du/ ’army’).[5] Although it may be possible to disambiguate whether a noun is a proper noun or not using orthographical cues such as initial capitalization and/or suffix separation characters, this may not always be possible and one may have to use again techniques akin to word sense disambiguation.
6. The problem above is further complicated in cases where a proper noun is stressed differently when it denotes a place than when it denotes person, e.g., *Aydın* (/”aj-d1n/ ’city’) vs. *Aydın* (/aj-”d1n/ ’person’). To disambiguate such cases one would have to resort to named-entity recognition techniques.

Since morphological disambiguation is one of the major pronunciation disambiguation tools, in the rest of the paper we present a morphological disambiguation scheme that is based on the concept of *distinguishing tags* – that subset of morphological feature tags in a morphological analysis that is sufficient to uniquely identify that analysis, and apply it to pronunciation disambiguation. The use of word sense disambiguation and named entity recognition techniques are outside the scope of this paper.

3 Morphological Disambiguation

Morphological analysis is a prior step to be performed in many natural language processing applications. A morphological analyzer produces all the possible morphological parses of an input word. An ambiguity arises if more than one analysis are generated for one word.[6] Given a sequence of words, selecting the correct analysis among alternatives for each is defined as morphological disambiguation.

Let $W = w_1, w_2, \ldots, w_n$ be a sequence of `n` words. The set of morphological parses of w_i will be denoted by $M_i = \{m_{i,1}, m_{i,2}, \ldots m_{i,a_i}\}$, where a_i denotes

[5] For example, *3. Ordu Futbol Şenliği* (Third Ordu Soccer Festival) vs. *3. Ordu Futbol Takımı* (Third Army Soccer Team).

[6] In a typical running Turkish text, every word has on the average close to 2 morphological interpretations but about 60% actually have a single interpretation.

Table 2. Possible morphological parses and pronunciations of the word `karın`

Morphological Analysis	*Pronunciation*	*English Translation*
`kar+Noun+A3sg+P2sg+Nom` (m_1)	/ca:-”r1n/ (s_1)	your profit
	/ka-”r1n/ (s_2)	your snow
`kar+Noun+A3sg+Pnon+Gen` (m_2)	/ca:-”r1n/ (s_1)	of the profit
	/ka-”r1n/ (s_2)	of the snow
`kar+Verb+Pos+Imp+A2sg` (m_3)	/”ka-r1n/ (s_3)	mix it
`karı+Noun+A3sg+P2sg+Nom` (m_4)	/ka-”r1n/ (s_2)	your wife
`karın+Noun+A3sg+Pnon+Nom` (m_5)	/ka-”r1n/ (s_2)	belly

number of distinct parses for w_i. Morphological disambiguation aims to select the correct $m_{i,j}$ for each w_i in the given context. Associated with each morphological analysis m_j, are one or more possible pronunciations of the word, s_k[7] under that morphological interpretation. For example, Table 2, shows all morphological parses of word `karın` in Turkish with the corresponding pronunciations in SAMPA format, and the English gloss. There are five distinct morphological analyses but only three different pronunciations. Pronunciations 1 and 2 are associated with morphological parses 1 and 2, since the sense of the word is not in the morphological analysis. If the morphological disambiguation process for an occurrence of the word `karın` in a context results in m_3, the pronunciation would be s_3. Otherwise, if m_4 or m_5 is selected, then the reading is s_2. On the other hand, word sense disambiguation would be required to select the pronunciation in the cases of m_1 or m_2.

3.1 Modeling with Distinguishing Tags

Turkish is an agglutinative language with a highly productive inflectional and derivational morphology. Morphosyntactic analyses of words in the language require large number of tags to indicate all the morphosyntactic and morphosemantic tags encoded in a word. Statistical disambiguation methods that rely on n-gram statistics of all the tags suffer from data sparseness problem, since the tags set is very large. Tür *et al.* [6] have split up the morphological analyses across any derivational boundaries into inflectional groups (IGs) to partially overcome the problem and proposed to model each morphological parse via these IGs, each of which is actually a shorter sequence of feature tags.

This work proposes to use distinguishing tag sets of the final IG and major POS of the first IG. The *distinguishing tag sets* (DTS) of a morphological parse are defined as follows: Given a morphological analysis, let α represent the set of all subsets of the features used in its final IG. The distinguishing tag sets of the analysis are the elements from α with the smallest size, such that if we determine that the correct analysis has those features, then we can uniquely identify the complete parse.

[7] We will avoid using multiples indices to denote morphological parses pronunciations and just refer to them with m_j and s_k respectively when the word index is obvious from the context.

For example, the word `çalışmaları` has the following parses:

1. `çalış+Verb+Pos^DB+Noun+Inf2+A3pl+P3sg+Nom`
2. `çalış+Verb+Pos^DB+Noun+Inf2+A3pl+Pnon+Acc`
3. `çalış+Verb+Pos^DB+Noun+Inf2+A3pl+P3pl+Nom`
4. `çalış+Verb+Pos^DB+Noun+Inf2+A3sg+P3pl+Nom`

The `^DB` symbols mark the derivation boundaries. Every morphological parse of the word is split up into inflectional groups (IG) across any derivation boundaries. Note that, if no derivation boundaries exist in an analysis, then its first and final IG is the same. The second column of below shows the distinguishing tag sets for each analysis of the word *barışmış*.

Parse	*Possible Sets of Distinguishing Tags*	*Major POS of first IG*
1	{ +P3sg }	+Verb
2	{ +Pnon } , { +Acc }	+Verb
3	{ +A3pl,+P3pl }	+Verb
4	{ +A3sg }	+Verb

So if we determine during disambiguation that the correct parse is {+P3sg} , then that is sufficient to deduce that the correct analysis is the first one. Similarly {+A3pl,+P3pl} and {+A3sg} imply the third and fourth parses. There are two distinct DTS for the second morphological analysis and either {+Pnon} or {+Acc} identifies it.

We model each $m_{i,j}$ by the DTS and the major part-of-speech tag[8] of the first IG. Note that, a morphological analysis may have more than one DTS, but only one root major POS. Following the notation used in statement of the problem, for each $m_{i,j}$, let $O_{i,j}$ be the major POS tag of the first IG and $DTS^{l}_{i,j}$ be one of its DTS. The set $R_{i,j} = \{(O_{i,j}, DTS^{1}_{i,j}), (O_{i,j}, DTS^{2}_{i,j}), \ldots, (O_{i,j}, DTS^{q}_{i,j})\}$ contains all distinct representations of $m_{i,j}$, assuming there are q different DTS identifying the analysis. For the example word w_i `çalışmaları`; $R_{i,1}$= {(Verb, P3sg)}, $R_{i,2}$= {(Verb, Pnon), (Verb, Acc)}, $R_{i,3}$= {(Verb, (A3pl,P3pl))}, and $R_{i,4}$= {(Verb, A3sg)}. Remembering that a_i was the number of distinct analyses of word w_i, $R_i = R_{i,1} \cup R_{i,1} \cup \ldots \cup R_{i,a_i}$ contains all representations of all morphological parses for w_i, where each element $(O_{i,j}, DTS^{1}_{i,j})$ of the set uniquely selects an analysis.

Finally, let us define t_i as the element selected from R_i by the morphological disambiguator. The $t_i \in R_i$ determines the morphological parse. If $t_i \in R_{i,1}$ then $m_{i,1}$ is selected; if $t_i \in R_{i,2}$ then $m_{i,2}$ is selected. Similarly, $m_{i,3}$ and $m_{i,4}$ are selected if $t_i \in R_{i,3}$ and $t_i \in R_{i,4}$ respectively.

[8] This study is performed on the outputs of the Turkish morphological analyzer [9], which assigns following tags as major POS to every IG: `+Noun`, `+Adj` (adjective), `+Adverb`, `+Verb`, `+Det` (determiner), `+Conj` (conjunction), `+Pron` (pronoun), `+Dup` (duplication), `+Interj` (interjection), `+Ques` (question), `+Postp` (postposition), `+Num` (number), `+Punc` (punctuation), `+BSTag` (beginning of sentence), `+ESTag` (end of sentence).

The disambiguation of a given sequence of words begins with the identification of $O_{i,j}$ and $DTS^{l}_{i,j}$ for all possible values of i,j, and l in the word sequence. A Hidden Markov Model is then constructed to compute for the sequence $T = t_1, t_2 \cdots t_i \cdots t_n$, where each t_i denotes a root POS and DTS combination referring to a unique morphological analysis $m_{i,j}$. The sequence T is computed in the usual way using Equation (1) by maximizing the probability $P(T \mid W)$:

$$\underset{T}{\text{argmax}}\ P(T \mid W) = \underset{T}{\text{argmax}}\ \frac{P(T) \times P(W \mid T)}{P(W)} = \underset{T}{\text{argmax}} P(T) \times P(W \mid T) \quad (1)$$

since $P(W)$ is constant for every selection of T.

Turkish does not have morphological generation ambiguity[9] so that the word formed by the given set of tags is unique. This implies $P(W \mid T) = 1$ all the time, and hence the equation simplifies to:

$$\underset{T}{\text{argmax}}\ P(T \mid W) = \underset{T}{\text{argmax}}\ P(T) \quad (2)$$

The trigram approximation for P(T)

$$P(T) = \prod_{i=1}^{n} P(t_i \mid t_{i-1}, t_{i-2}) \quad (3)$$

can now be written as

$$P(T) = \prod_{i=1}^{n} P(\ (O_{i,x}, DTS^{l_i}_{i,x}) \mid (O_{i-1,y}, DTS^{l_{i-1}}_{i-1,y}),\ (O_{i-2,z}, DTS^{l_{i-2}}_{i-2,z})\ ) \quad (4)$$

where $t_i \in R_i$, $t_{i-1} \in R_{i-1}$, and $t_{i-2} \in R_{i-2}$; and x, y, z range over the respective number of ambiguous parses – $1 \leq x \leq a_i$, $1 \leq y \leq a_{i-1}$, $1 \leq z \leq a_{i-2}$.

The Viterbi algorithm can now be used on the expanded trigram model to find the highest scoring path, which makes up the sequence T. As an example, Figure 1 illustrates the distinguishing tag modeling used in the morphological disambiguation of the utterance `"Sadece doktora çalışmaları tartışıldı."` (*Only the Ph.D. studies were discussed.*). The morphological parses of the words along with their ($O_{i,j}$,$DTS^{l}_{i,j}$) pairs are as:

1. `sadece+Adverb` : $t_{1,1}$=(Adverb,Adverb)
   `sadece+Adj^DB+Adj+Asif` : $t_{1,2}$=(Adj,Adj)[10]
2. `doktor+Noun+A3sg+Pnon+Dat` : $t_{2,1}$=(Noun,Dat)
   `doktora+Noun+A3sg+Pnon+Nom` : $t_{2,2}$=(Noun,Nom)
3. `çalışmaları` : $t_{3,1}$, $t_{3,2}$, $t_{3,3}$, $t_{3,4}$, and $t_{3,5}$ were given in the previous example.
4. `tartış+Verb^DB+Verb+Pass+Pos+Past+A3sg` : $t_{4,1}$=(Verb,Verb), $t_{4,2}$=(Verb,Pass), $t_{4,3}$=(Verb,Pos), $t_{4,4}$=(Verb,Past), $t_{4,5}$=(Verb,A3sg)

[9] Refer to Tür *et al.* [6] for a few rarely seen words that have this ambiguity.

[10] `+Asif` is a semantic marker and thus not included in the DTS generation.

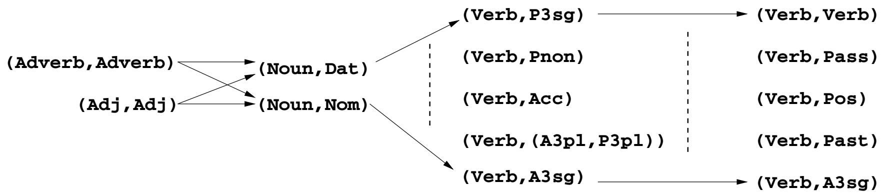

Fig. 1. Sample sentence modeled with distinguishing tags

It should be noted that we can also use a formulation using *argmin* instead of *argmax* to find the worst scoring parses and remove them to *reduce* morphological ambiguity. This may also be meaningful in pronunciation disambiguation that does not need full-fledged morphological disambiguation.

4 Implementation and Results

We have used the same 1 million-word corpus used by Tür *et al.* [6]. The morphological analysis of each word was generated by the Turkish morphological analyzer [9]. The current analyzer produces 116 feature tags of which 28 are used to label the semantic features. Throughout the study those semantic tags were discarded and our model considered only the inflectional tags while computing the DTS. If two or more parses of a word differ only in semantic tags, then they are assumed to be the same analysis.

Before any disambiguation, the precision was 55.9% and the morphological ambiguity was 1.8 parses per word. A preprocessing step which does not reduce the recall, was run on the test set before the actual disambiguator is run. The preprocessor performs some rule-based reductions to eliminate analyses that can only be seen in very restricted domains or are very infrequent or obsolete root words. At this stage the morphological ambiguity becomes 1.45 while precision rises to 68%.

The morphological parses selected by the disambiguator by Tür *et al.* ([6]) were used to train the statistical model using of the CMU-Cambridge statistical language modeling toolkit[10]. The corpus was split into 10 approximately equal pieces and a 10-fold cross validation scheme was used. In each of the tests, 9 of the segments were used as the training set from which the trigram statistics were trained, and one segment was used as the test set.

No disambiguation was performed on the ambiguous proper names because selecting their correct pronunciations require special processing beyond the morphosyntactic analysis as described in Section 2. They were left with all their morphosyntactic parses and corresponding pronunciations. Although this lowers precision a little bit, it lets further named entity recognition tasks to be performed on the corpus without loss of recall.

As described earlier, there are two choices for the disambiguation process. The first is to run the model with *argmax* and select the best scoring tag sequence and the second is to run with *argmin* to select the worst scoring tag sequence

Table 3. Morphological and pronunciations disambiguation performance with select-best and throw-worst approaches. P,R, and A represent *precision*, *recall* and *ambiguity* respectively.

Fold	SELECT-BEST						THROW-WORST					
	Morp. Dis.			Pron. Dis.			Morp. Dis.			Pron. Dis.		
ID	% of P	% of R	A	% of P	% of R	A	% of P	% of R	A	% of P	% of R	A
1	84.97	92.80	1.09	98.28	99.13	1.01	80.88	96.05	1.18	98.06	99.58	1.01
2	84.20	92.30	1.09	98.29	99.10	1.01	80.06	95.86	1.20	97.96	99.53	1.02
3	84.26	92.81	1.10	98.23	99.17	1.01	80.36	95.95	1.19	97.96	99.55	1.02
4	84.95	92.93	1.09	98.28	99.09	1.01	80.88	95.89	1.18	98.02	99.55	1.01
5	84.24	92.87	1.10	98.04	99.16	1.01	80.37	96.05	1.19	97.72	99.54	1.01
6	84.66	92.90	1.09	98.11	99.10	1.01	80.74	96.02	1.19	97.75	99.49	1.02
7	83.91	92.97	1.11	98.13	99.25	1.01	80.13	95.97	1.20	97.84	99.60	1.02
8	84.21	92.99	1.10	98.12	99.22	1.01	80.40	96.02	1.19	98.70	99.58	1.02
9	84.15	93.04	1.10	97.90	99.11	1.01	80.60	96.10	1.19	97.64	99.57	1.02
10	82.60	90.80	1.09	98.11	99.01	1.01	77.97	92.99	1.19	97.81	99.45	1.02
AVG	**84.22**	**92.64**	**1.10**	**98.15**	**99.13**	**1.01**	**80.24**	**95.69**	**1.19**	**97.95**	**99.54**	**1.02**

and discard the parses on that sequence away (unless the only parse for a word). The results of each experiment are summarized in Table 3 that lists both the morphological disambiguation and pronunciation disambiguation results.[11] The evaluations are done by *precision*, *recall* and *ambiguity* metrics.

5 Conclusions

We presented our results for pronunciation disambiguation in Turkish using only restricted morphological disambiguation. The morphosyntactic disambiguator was based on distinguishing tag sets which was formulated as a solution to the data sparseness problem. The number of distinct distinguishing tag sets used in the study is 374. Note that Tür *et al.* [6] reported 2194 distinct inflectional groups on the same corpus. Using a 10-fold validation experiment we found that with the throw-worst approach (using *argmin*), our approach disambiguates the pronunciations with 99.54% recall and 97.95% precision. Although this approach does not fully disambiguate the morphological ambiguities, it performs well on the disambiguation of pronunciations. The more aggressive select-best approach (using *argmax*) gives better precision values, but the recall gets lower. The remaining pronunciation ambiguities need additional techniques such as named entity recognition and word sense disambiguation.

[11] In Table 3, note that A is a little bit higher than its expected exact value of 1 in select-best strategy. The reason is that although the analyses of a word that differ only in their semantic tags are considered to be the same for disambiguation purposes as mentioned in Section 4, they are counted separately while extracting the statistics.

References

1. Oflazer, K., Inkelas, S.: The architecture and the implementation of a finite state pronunciation lexicon for Turkish. To appear in Computer Speech and Language (2005)
2. Brill, E.: A simple rule-based part-of-speech tagger. In: Proceedings of the third Applied Natural Language Processing, Trento, Italy (1992)
3. Oflazer, K., Tür, G.: Combining hand-crafted rules and unsupervised learning in constraint-based morphological disambiguation. In Brill, E., Church, K., eds.: Proceedings of the ACL-SIGDAT Conference on Empirical Methods in Natural Language Processing. (1996)
4. Ezeiza, N., Alegria, I., Arriola, J., Urizar, R., Aduriz, I.: Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada (1998) 379–384
5. Hajic, J., Krbec, P., Kveton, P., Oliva, K., Petkevic, V.: Serial combination of rules and statistics: A case study in Czech tagging. In: Proceedings of ACL'01, Toulouse, France (2001)
6. Tür, D.Z.H., Oflazer, K., Tür, G.: Statistical morphological disambiguation for agglutinative languages. Computers and the Humanities **36** (2002) 381–410
7. Oflazer, K., Kuruöz, I.: Tagging and morphological disambiguation of Turkish text. In: Proceedings of the 4th Applied Natural Language Processing Conference, ACL (1994) 144–149
8. Oflazer, K., Tür, G.: Morphological disambiguation by voting constraints. In: Proceedings of ACL'97, Madrid, Spain (1997)
9. Oflazer, K.: Two level description of Turkish morphology. Literary and Linguistic Computing **9** (1994) 137–148
10. Clarkson, P., Rosenfeld, R.: Statistical language modeling using the CMU-Cambridge toolkit. In: Proceedings of Eurospeech'97, Rhodes, Greece (1997) 2707–2710