Chapter 4
Language Modeling for Turkish Text and Speech Processing

Ebru Arısoy and Murat Saraçlar

Abstract This chapter presents an overview of language modeling followed by a discussion of the challenges in Turkish language modeling. Sub-lexical units are commonly used to reduce the high out-of-vocabulary (OOV) rates of morphologically rich languages. These units are either obtained by morphological analysis or by unsupervised statistical techniques. For Turkish, the morphological analysis yields word segmentations both at the lexical and surface forms which can be used as sub-lexical language modeling units. Discriminative language models, which outperform generative models for various tasks, allow for easy integration of morphological and syntactic features into language modeling. The chapter provides a review of both generative and discriminative approaches for Turkish language modeling.

4.1 Introduction

A statistical language model assigns a probability distribution over all possible word strings in a language. The ultimate goal in statistical language modeling is to find probability estimates for word strings that are as close as possible to their true distribution. In the last couple of decades, a number of statistical techniques have been proposed to appropriately model natural languages. These techniques employ large amounts of text data to robustly estimate model parameters which are then used to estimate probabilities of unseen text.

Statistical language models are used in many natural language applications such as speech recognition, statistical machine translation, handwriting recognition, and

E. Arısoy
MEF University, Istanbul, Turkey
e-mail: ebruarisoy.saraclar@mef.edu.tr

M. Saraçlar (✉)
Boğaziçi University, Istanbul, Turkey
e-mail: murat.saraclar@boun.edu.tr

K. Oflazer, M. Saraçlar (eds.), *Turkish Natural Language Processing*,
Theory and Applications of Natural Language Processing,
https://doi.org/10.1007/978-3-319-90165-7_4

spelling correction, as a crucial component to improve the performance of these applications. In these and other similar applications, statistical language models provide prior probability estimates and play the role of the source model in communication theory inspired source-channel formulations of such applications. A typical formulation of these applications allows language models to be used as a predictor that can assign a probability estimate to the next word given the contextual history. Some applications employ more complex language models in reranking scenarios where alternative hypotheses generated by a simpler system are rescored or reranked using additional information. A typical example is the feature-based discriminative language model where model parameters associated with many overlapping features are used to define a cost or conditional probability of the word sequences. Such a model then enables the selection of the best hypothesis among the alternatives based on the scores assigned by the model.

This chapter focuses on language modeling mainly for Turkish text and speech processing applications. First we introduce the foundations of language modeling and describe the popular approaches to language modeling, then we explain the challenges that Turkish presents for language modeling. After reviewing various techniques proposed for morphologically rich languages including Turkish, we summarize the approaches used for Turkish language modeling.

4.2 Language Modeling

Statistical language models assign a prior probability, $P(W)$, to every word string $W = w_1\ w_2\ \ldots\ w_N$ in a language. Using the chain rule, the prior probability of a word string can be decomposed into the following form:

$$P(W) = P(w_1\ w_2\ \ldots\ w_N) = \prod_{k=1}^{N} P(w_k|w_1\ \ldots\ w_{k-1}). \tag{4.1}$$

Here the prior probability is calculated in terms of the dependencies of words to a group of preceding words, $w_1\ \ldots\ w_{k-1}$, called the "history." These conditional probabilities need to be estimated in order to determine $P(W)$. It is, however, not practical to obtain the prior probability as given in Eq. (4.1) for two main reasons. First, if the history is too long, it is not possible to robustly estimate the conditional probabilities, $P(w_k|w_1\ \ldots\ w_{k-1})$. Second, it is not entirely true that the probability of a word depends on *all* the words in its entire history. It is more practical and realistic to assign histories to equivalence classes $\Psi(w_1 \ldots\ w_{k-1})$ (Jelinek 1997). Equivalence classes change Eq. (4.1) into the following form:

$$P(W) = P(w_1\ w_2\ \ldots\ w_N) = \prod_{k=1}^{N} P(w_k|\Psi(w_1\ \ldots\ w_{k-1})) \tag{4.2}$$

While the equivalence classes can be based on any classification of the words in the history, or their syntactic and semantic information, the most common approach is based on a very simple equivalence classification which utilizes only the $n-1$ preceding words as the history. This approach results in the widely used n-gram language models, and $P(W)$ is approximated as

$$P(W) = P(w_1\ w_2\ \ldots\ w_N) \approx \prod_{k=1}^{N} P(w_k|w_{k-n+1}\ldots\ w_{k-1}) \tag{4.3}$$

The n-gram language model probabilities are estimated from a text corpus related to the application domain with Maximum Likelihood Estimation (MLE). In other words, n-gram probabilities are estimated by counting the occurrences of a particular n-gram in the text data and dividing this count by the number of occurrences of all n-grams that start with the same sequence of $n-1$ words:

$$P(w_k|w_{k-n+1}\ \ldots\ w_{k-1}) = \frac{C(w_{k-n+1}\ \ldots\ w_{k-1}\ w_k)}{C(w_{k-n+1}\ \ldots\ w_{k-1})} \tag{4.4}$$

where $C(\cdot)$ represents the number of occurrences of the word string given in parentheses in the text data.

If the language model vocabulary contains $|V|$ words, then there may be up to $|V|^n$ n-gram probabilities to be calculated—thus higher order n-grams need a much larger set of language model parameters. Robust estimation of n-gram probabilities with MLE critically depends on the availability of large amounts of text data. However experience with many applications has shown that 3/4/5-gram models are quite satisfactory and higher order models do not provide any further benefits.

The quality of the statistical language models can be best evaluated using the performance of the applications they are used in—for example, speech recognition or statistical machine translation. An alternative approach without including the overall system into the evaluation is to rely on *perplexity* to gauge the generalization capacity of the proposed language model on a separate text that is not seen during model training. Formally, perplexity is defined as:

$$PP(w_1, w_2, \cdots, w_N) = 2^{-\frac{1}{N}\log_2 P(w_1, w_2, \cdots, w_N)} \tag{4.5}$$

In other words, perplexity shows us how well a language model trained on a text data does on an unseen text data. Minimizing the perplexity corresponds to maximizing the probability of the test data. Even though a lower perplexity usually means a better language model with more accurate prediction performance, perplexity may not always be directly correlated with application performance.

One of the problems in n-gram language modeling is data sparseness. Any finite training corpus contains only a subset of all possible n-grams. So, MLE will assign zero probability to all *unseen* n-grams. A test sentence containing such n-grams not seen in the training corpus will also be assigned zero probability according to

Eq. (4.3). In order to prevent this, a technique known as *smoothing* is employed to reserve some of the probability mass to unseen n-grams so that no n-gram gets zero probability. This also means that this mass comes from the probabilities of the observed n-grams leading to slight reductions in their probabilities. Smoothing techniques thus lead to better language model estimates for unseen data.

Interpolation and back-off smoothing are the most common smoothing methods. In interpolation, higher and lower order n-gram models are linearly interpolated. In back-off smoothing, when a higher order n-gram model assigns zero probability to a particular n-gram, the model backs off to a lower order n-gram model. Good-Turing, Katz, and Kneser-Ney are some examples of popular smoothing algorithms. See Chen and Goodman (1999) for a survey of smoothing approaches for statistical language models.

In addition to these smoothing techniques, class-based n-gram language models (Brown et al. 1992) and continuous space language models (Bengio et al. 2003; Schwenk 2007) have been used to estimate unseen event probabilities more robustly. These approaches try to make more reasonable predictions for the unseen histories by assuming that they are similar to the histories that have been seen in the training data. Class-based language models group words into classes, while continuous space language models project words into a higher dimensional continuous space, with the expectation that words that are semantically or grammatically related will be grouped into the same class or mapped to similar locations in the continuous space. The main goal of these models is to generalize well to unseen n-grams.

One drawback of the conventional n-gram language models is their reliance on only the last $n-1$ words in the history. However, there are many additional sources of information, such as morphology, syntax, and semantics, that can be useful while predicting the probability of the next word. Such additional linguistic information can be either incorporated into the history of the n-gram models or encoded as a set of features to be utilized in feature-based language models.

Structured language models (Chelba and Jelinek 2000), probabilistic top-down parsing in language modeling (Roark 2001), and Super ARV language models (Wang and Harper 2002) are some example approaches that incorporate syntactic information into the n-gram history. The factored language model (Bilmes and Kirchhoff 2003) is another example that incorporates syntactic as well as morphological information into the n-gram history.

Feature-based models allow for easy integration of arbitrary knowledge sources into language modeling by encoding relevant information as a set of features. The maximum entropy language model (Rosenfeld 1994) is a popular example of this type, where the conditional probabilities are calculated with an exponential model,

$$P(w|h) = \frac{1}{Z(h)} e^{\sum_i \alpha_i \Phi_i(h,w)}. \tag{4.6}$$

Here, $Z(h)$ is a normalization term and $\Phi_i(h, w)$'s are arbitrary features which are functions of the word w and the history h. The whole sentence maximum entropy model (Rosenfeld et al. 2001) assigns a probability to the whole sentence using the

features $\Phi_i(W)$ with a constant normalization term Z:

$$P(W) = \frac{1}{Z} e^{\sum_i \alpha_i \Phi_i(W)}. \tag{4.7}$$

Discriminative language models (DLMs) (Roark et al. 2007) have been proposed as a complementary approach to the state-of-the-art n-gram language modeling. There are mainly two advantages of DLMs over n-grams. The first advantage is improved parameter estimation with discriminative training, since DLMs utilize both positive and negative examples to optimize an objective function that is directly related with the system performance. In training a DLM, positive examples are the correct or meaningful sentences in a language while negative examples are word sequences that are not legitimate or meaningful sentences in the language.

The second advantage is the ease of incorporating many information sources such as morphology, syntax, and semantics into language modeling. As a result, DLMs have been demonstrated to outperform generative n-gram language models. Linear and log-linear models have been successfully applied to discriminative language modeling for speech recognition (Roark et al. 2004, 2007; Collins et al. 2005). In DLMs based on linear models, model parameters are used to define a cost, $F(W)$, on the word sequence

$$F(W) = \sum_i \alpha_i \Phi_i(W). \tag{4.8}$$

In DLMs based on log-linear models, the cost $F(W)$ has the same form as the log of the probability given by the whole sentence maximum entropy model

$$F(W) = \sum_i \alpha_i \Phi_i(W) - \log Z, \tag{4.9}$$

where Z is approximated by summing over the alternative hypotheses. The details of the DLM framework will be given in Sect. 4.6.

4.3 Challenges in Statistical Language Modeling for Turkish

In the context of language modeling, two aspects of Turkish, very productive agglutinative morphology leading to a very large vocabulary, and free constituent order make statistical language modeling rather challenging, especially for applications such as automatic speech recognition (ASR) and statistical machine translation (SMT).

State-of-the-art ASR and SMT systems utilize predetermined and finite vocabularies that contain the most frequent words related to the application domain. The words that do not occur in the vocabulary but are encountered by the ASR or SMT

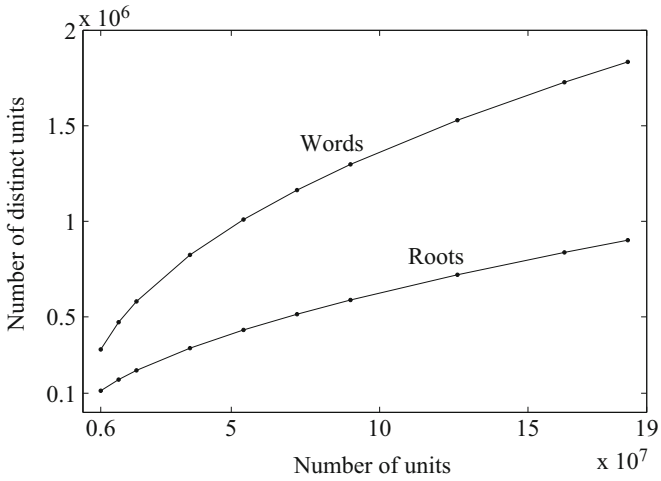


Fig. 4.1 Vocabulary growth curves for words and roots

system are called Out-Of-Vocabulary (OOV) words. Existence of OOV words is one of the causes of degradation in system performance. For instance in an ASR system, if a word is not in the vocabulary and it is uttered by a speaker, it has no chance to be recognized correctly. Hetherington (1995) estimates that as a rule of thumb an OOV word leads to on the average 1.5 recognition errors.

As described in earlier chapters, the very productive morphology of Turkish yields many unique word forms, making it difficult to have a fixed vocabulary covering all these words. Figure 4.1 illustrates the growth for unique Turkish words and roots as a function of the number of tokens in a text corpus of 182.3M word tokens (units) and 1.8M word types (distinct units). It can be observed that the increase in the number of distinct units with the increasing amount of data is much higher for words compared to roots which is an expected result for Turkish. From the morphological analysis of these Turkish words, we have also observed that on the average each root generates 204 words and each word is composed of on the average 1.7 morphemes including the root.[1] The verb *etmek* "to do" accounts for 3348 unique words—the maximum number for any of the roots. The word form *ruhsat+lan+dır+ıl+ama+ma+sı+nda+ki* is an example with the maximum number of morphemes but only occurs once in the corpus.

This significant word vocabulary growth results in high OOV rates even for vocabulary sizes that would be considered as large for English. Figure 4.2 shows the OOV rates calculated on a test data of 23K words, for different vocabulary sizes. For instance, around 9% OOV rate is achieved with a vocabulary size of 60K words.

[1]But as noted in Chap. 2, most high-frequency words have a single morpheme so most likely inflected words have more than 1.7 morphemes.

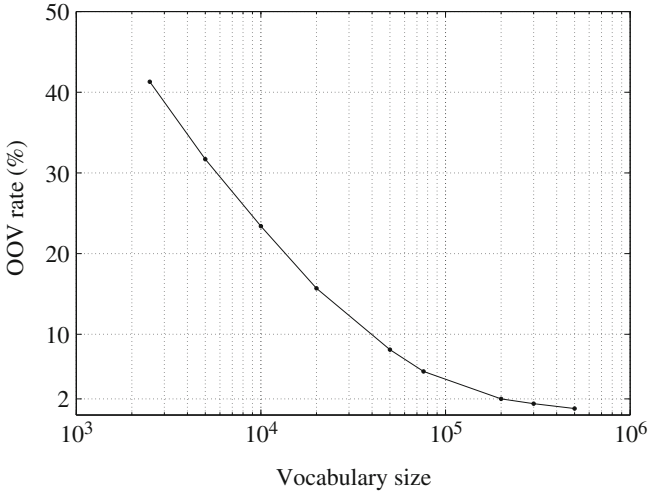


Fig. 4.2 OOV rates for Turkish with different vocabulary sizes

However, with an optimized 60K word lexicon for English, the OOV rate is less than 1% for North American business news text (Rosenfeld 1995). Other morphologically rich languages such as Finnish, Estonian, Hungarian, and Czech also suffer from high OOV rates: 15% OOV with a 69K lexicon for Finnish (Hirsimäki et al. 2006), 10% OOV with a 60K lexicon for Estonian (Kurimo et al. 2006), 15% OOV with a 20K lexicon for Hungarian (Mihajlik et al. 2007), and 8.3% OOV with a 60K lexicon for Czech (Podvesky and Machek 2005). Even though these numbers are not directly comparable with each other, they indicate that high OOV rates are a major problem for morphologically rich languages. Therefore, addressing the OOV problem is crucial for the performance of downstream applications systems that make use of statistical language models.

The free word order is another challenge for statistical language modeling. The relatively free word order contributes to the sparseness data and this can lead to non-robust n-gram language model estimates. However this is more of a problem for speech recognition applications or processing of informal texts—in formal text such as news the dominant constituent order is subject-object-verb but there are no reliable statistics on the distribution of different constituent order in large Turkish corpora. We will not be addressing this issue in the rest of this chapter.

4.4 Sub-lexical Units for Statistical Language Modeling

A commonly proposed solution for reducing high OOV rates for morphologically rich languages is to use sub-lexical units for language modeling. In sub-lexical language modeling, the vocabulary is composed of sub-lexical units instead of

words. These could be letters, syllables, morphemes or combination of morphemes or arbitrary word segments. In order to address the OOV problem, the sub-lexical unit vocabulary should be capable of covering most of the words of a language, and clearly these sub-lexical units should be meaningful for prediction using language models. They should have limited confusion and avoid over-generation. For instance, if the letters are used as sub-lexical units, only a vocabulary of 29 letters of the Turkish alphabet will cover all the words in the language. However, letters are not logical sub-lexical unit choices since they require very long histories for accurate language model predictions and they allow more confusable choices in, for instance, speech recognition. Note also that the perplexities of language models based on different units are not directly comparable due to each model having different OOV rates and different number of tokens for evaluation. Assuming no OOVs, perplexity of sub-lexical language models need to be normalized by the number of word tokens for a fair comparison. However, a better way of comparing sub-lexical and word language models is directly measuring the task performance.

Sub-lexical units can be classified as being linguistic or statistical, based on the underlying algorithm utilized in segmenting words into sub-lexical units. Linguistic sub-lexical units are obtained with rule-based morphological analyzers while statistical sub-lexical units are obtained with statistical segmentation approaches that rely on unsupervised model of word segmentation.

4.4.1 Linguistic Sub-lexical Units

In agglutinative languages like Turkish, words are formed as a concatenation of stems and affixes. Therefore, linguistic units such as stems, affixes, or their groupings can be considered as natural choices of sub-lexical units for language modeling. In language modeling with linguistic sub-lexical units, the words are split into morphemes using morphological analyzers, and then a vocabulary composed of chosen morphological units is built for language modeling. However, there is a trade off between using long and short units: long units, e.g., full words will result in OOV problem while shorter units (e.g., morphemes) will require larger n-grams for prediction and risk assigning probabilities to non-words because of over-generation. Since morphemes might be very short, as short as a single letter, Kanevsky et al. (1998) have suggested using stems and endings as vocabulary units as a compromise between words and morphemes, where an ending is what is left after removing the root from the word.

Morphemes, stems, and endings are examples of commonly used linguistic sub-lexical units in language modeling of agglutinative languages like Turkish, Korean, Finnish, Estonian, and Hungarian and highly inflectional languages like Czech, Slovenian, and Arabic. Morpheme-based language models were utilized for language modeling of Korean, another agglutinative language, and to deal with the coarticulation problem rising from very short morphemes, frequent and short morpheme pairs were merged before modeling (Kwon and Park 2003).

Morpheme-based language models were also investigated for language modeling of Finnish (Hirsimäki et al. 2006), Estonian (Alumäe 2006), and Hungarian (Mihajlik et al. 2007), all also agglutinative. These researchers also compared linguistic sub-lexical units with their statistical counterparts for ASR. Kirchhoff et al. (2006) and Choueiter et al. (2006) applied morphology-based language modeling to Arabic ASR and reported better recognition results than words. Rotovnik et al. (2007) used stem and endings for Slovenian language modeling for ASR. Additional constraints to the ASR decoder, such as restricting the correct stem and ending order, and limiting the number of endings for an individual stem were found to reduce over-generation.

The main disadvantage of linguistic sub-lexical units is the need for expert knowledge of the language for building the morphological analyzers. Thus they are not applicable to languages lacking such morphological tools. Additionally, even if a morphological analyzer is available, usually a fixed limited root vocabulary may not necessarily help with the OOV problem. For instance, a Turkish morphological analyzer (Sak et al. 2011) with 54.3K roots can analyze only 52.2% of the 2.2M unique words in a text corpus of 212M words. However, the words that the morphological analyzer cannot parse are usually rare words and only account for about 3% of the word tokens in the text corpus. Hence, this limitation may not necessarily have much impact on the statistical language model. A more important concern is the need for morphological disambiguation of multiple analyses of words.

4.4.2 Statistical Sub-lexical Units

Statistical sub-lexical units are morpheme-like units or segments obtained by data driven approaches, usually in an unsupervised manner. The main advantage of statistical sub-lexical units is that they do not rely on a manually constructed morphological analyzer. These segments do not necessarily match with the linguistic morphemes, however, they are "meaningful" units in terms of language modeling.

One of the earliest works in this area, Harris (1967) posited morpheme boundaries in a word by using letter transition frequencies with the assumption that the predictability of a letter will decrease at the morpheme boundaries.

The last 15 years have seen a surge in data-driven algorithms for unsupervised morpheme discovery based on probabilistic models as well as some heuristics. One of the algorithms with publicly available software is Linguistica (Goldsmith 2001) that utilizes the minimum description length (MDL) principle to learn morphological segmentations in an unsupervised way, aiming to find the segmentations as close as possible to the true morphemes. Whittaker and Woodland (2000), motivated by the productive morphology of Russian, aim to obtain sub-lexical units (called particles) that maximize the likelihood of the training data using a bigram particle language model. In contrast to Linguistica, their algorithm does not aim to find the true morphological segmentations but instead searches for meaningful units for language modeling. Creutz and Lagus (2002, 2005) present

Morfessor, another algorithm for unsupervised discovery of morphemes. Morfessor was inspired by earlier work by Brent (1999) that explored word discovery during language acquisition of young children. Brent (1999) proposed a probabilistic model based on the MDL principle to recover word boundaries in a natural raw text from which they have been removed. The Morfessor algorithm also utilizes the MDL principle while learning a representation of the language in the data, as well as the most accurate segmentations. It is better suited for highly inflectional and agglutinative languages than Linguistica as it is designed to deal with languages with concatenative morphology. The annual Morpho Challenge competitions,[2] held since 2005, have helped the development of new algorithms for sub-lexical units. The Morfessor algorithm itself has been used as the baseline statistical sub-lexical approach in Morpho Challenge tracks and several different algorithms have competed against it.

Statistical sub-lexical units have been explored in language modeling of highly inflected and agglutinative languages. Hirsimäki et al. (2006), Kurimo et al. (2006), Siivola et al. (2003) applied Morfessor to Finnish, while Kurimo et al. (2006) applied it to Estonian and Mihajlik et al. (2007) to Hungarian. The performance of morpheme-based language models was compared with the language models built with Morfessor segmentations for Finnish (Hirsimäki et al. 2006) and Hungarian (Mihajlik et al. 2007) in the context of ASR. In Finnish ASR experiments, statistical units outperformed linguistic morphemes in news reading task where the number of foreign words that could not be handled by the morphological analyzer was quite high. In Hungarian ASR experiments for spontaneous speech, the best result was obtained with statistical segmentations. Hirsimäki (2009) describes the advances in building efficient speech recognition systems with Morfessor based segmentations. Kneissler and Klakow (2001) used an optimized sub-lexical approach for Finnish dictation and German street names recognition tasks. Pellegrini and Lamel (2007, 2009) modified the Morfessor algorithm to incorporate basic phonetic knowledge and explored its use for ASR of Amharic, a highly inflectional language.

4.5 Statistical Language Modeling for Turkish

This section reviews statistical language modeling units explored in Turkish text and speech processing systems. Figure 4.3 illustrates segmentations of the same Turkish phrase using different sub-lexical units. When applicable, the examples also show the lexical and surface form representations and "morphs" denote statistical sub-lexical units. In the rest of this section we will describe the details of Turkish language models based on these units.

[2] Aalto University, Finland. Department of Computer Science. "Morpho Challenge": morpho.aalto.fi/events/morphochallenge/ (Accessed Sept. 14, 2017).

Words: derneklerinin öncülüğünde

Syllables: der –nek –le –ri –nin ön –cü –lü –ğün –de

Morphemes
Lexical: dernek +lArH +nHn öncü +lHk +sH +nDA
Surface: dernek +leri +nin öncü +lüğ +ü +nde

Stem+Endings
Lexical: dernek +lArH+nHn öncü +lHk+sH+nDA
Surface: dernek +lerinin öncü +lüğünde

Morphs: dernek +lerinin öncü +lüğü +nde

Fig. 4.3 A Turkish phrase segmented into linguistic and statistical sub-lexical units

4.5.1 Language Modeling with Linguistic Sub-lexical Units

Over the last 15 years, various linguistic sub-lexical units for Turkish language modeling have been explored in the literature. Here we first review some of the earlier work and then summarize our work using such units.

Çarkı et al. (2000) were the first to investigate sub-lexical language models for Turkish. Due to the ambiguity in morphological analyses, they utilized syllables instead of morphemes as language modeling units and syllables were merged to obtain longer units with word-positioned syllable classes. While this approach addressed the serious OOV problem, it did not yield any improvements over the word-based language model built with a 30K vocabulary. Hakkani-Tür (2000) proposed groupings of morphemes, called inflectional groups as language modeling units. Mengüşoğlu and Deroo (2001) explored an extension of inflectional groups to n-gram language modeling as well as utilizing stem+ending models for Turkish. Dutağacı (2002) presented a comparative study of morpheme, stem+ending, and syllable language models in terms of generalization capacity of language models and OOV handling. ASR experiment results were also reported for these sub-lexical units, however, for a small vocabulary isolated word recognition task. This work was extended to continuous speech recognition by Arısoy (2004), and Arısoy et al. (2006) with a new model utilizing words, stem+endings and morphemes together in the same model vocabulary. Such a hybrid vocabulary combined model slightly outperformed the word model in terms of recognition accuracy when 10K units were used in combined and word bigram language models. Çiloğlu et al. (2004) compared bigram stem+ending model with a bigram stem model in terms of recognition accuracy in a small vocabulary ASR task and found that the stem model outperformed the stem+ending model when the language models were trained on

a very small text corpus (less than 1M words). However, the stem+ending model was shown to outperform stem model when the text corpus size was increased to approximately 6M words (Bayer et al. 2006).

Erdoğan et al. (2005) was one of the most comprehensive previous research on language modeling for Turkish ASR. The acoustic and language models in this work were trained on much larger amounts of data (34 h of speech corpus and 81M words text corpus). They investigated words, stem+endings and syllables as language modeling units and compared their performances on an ASR task and reported that the stem+ending model outperformed word and syllable models in recognition accuracy. This work also dealt with the over-generation problem of sub-lexical units by a post-processing approach imposing phonological constraints of Turkish and achieved further improvements over the best scoring stem+ending model.

Arısoy and Saraçlar (2009) presented another approach for dealing with the over-generation problem of sub-lexical units, especially for statistical sub-lexical units. This work along with Arısoy et al. (2009a) used a 200 million word text corpus collected from the web. The Turkish morphological parser described in Sak et al. (2011) was used to decompose words into morphemes and the Turkish morphological disambiguation tool developed by Sak et al. (2007) was used to disambiguate multiple morphological parses. Both the lexical and surface form representations of morphemes, stems and endings were used as linguistic sub-lexical units for Turkish. The details of these units are given in the following sections.

4.5.1.1 Surface Form Stem+Ending Model

Instead of using words as vocabulary items as in the word-based model, the surface form stem+ending model uses a vocabulary comprising surface form stem and endings and the words in the text data are split into their stems and endings. This is done by first extracting the stem from morphological analyses and taking the remaining part of the word as the ending.

In this approach, no morphological disambiguation was done. Instead Arısoy et al. (2009a) investigated building language models with all the ambiguous parses, with the parses with the smallest number of morphemes, and with randomly selected parses for each word token and type. They found no significant difference between the first two methods and these fared better than random choice of a parse. Sak et al. (2010) showed that utilizing the parse with the smallest number of morphemes performed slightly better than using the disambiguated parse in Turkish ASR. The method of selecting the parse with the smallest number of morphemes is not only extremely simple but also avoids more complex and error-prone approaches such as morphological disambiguation.

4.5.1.2 Lexical Form Stem+Ending Model

Morpholexical language models are trained as standard n-gram language models over morpholexical units. The one important advantage of using morpholexical units is that they allow conflating different surface forms of morphemes to one underlying form thereby alleviating the sparseness problem. For instance, the plural in Turkish is indicated by surface morphemes *+ler* or *+lar*, depending on the phonological (and not morphological) context. Thus representing these morphemes with a single lexical morpheme *+lAr* allows counts to be combined leading to more robust parameter estimation. Combining lexical morphemes also naturally leads to lexical stem+ending models (Arısoy et al. 2007).

Morpholexical language models have the advantage that they give probability estimates for sequences consisting of only valid words, that is they do not over-generate like the other sub-lexical models. Sak et al. (2012) have demonstrated the importance of both morphotactics and morphological disambiguation when producing the morpholexical units used for language modeling.

4.5.2 Statistical Sub-lexical Units: Morphs

As discussed earlier, statistical sub-lexical units obtained via unsupervised word segmentation algorithms have been used as an alternative to linguistic sub-lexical units. In fact, Turkish has been a part of the Morpho Challenge since 2007.[3]

Hacıoğlu et al. (2003) were the first to model Turkish with statistical sub-lexical units obtained with the Morfessor algorithm and showed that they outperform a word-based model with 60K word vocabulary, even though the language models were built on a text corpus containing only 2M words. Arısoy et al. (2009a) used statistical sub-lexical units for extensive experimentation using large corpora for Turkish ASR.

Arısoy et al. (2009b) proposed an enhanced Morfessor algorithm with phonetic features for Turkish. The main idea in this work was to incorporate simple phonetic knowledge of Turkish into Morfessor in order to improve the segmentations. Two main modifications were made to enhance Morfessor: a phone-based feature, called "DF" for distinctive feature, and a constraint called 'Cc' for confusion constraint. DF was directly incorporated into Morfessor's probability estimates and Cc was indirectly incorporated into Morfessor as a yes/no decision in accepting candidate splits. Both of these modifications aimed at reducing the number of confusable morphs in the segmentations by taking phonetic and syllable confusability into account.

[3]Aalto University, Finland. Department of Computer Science. "Morpho Challenge: Results": morpho.aalto.fi/events/morphochallenge/results-tur.html (Accessed Sept. 14, 2017).

4.6 Discriminative Language Modeling for Turkish

Recent ASR and MT systems utilize discriminative training methods on top of traditional generative models. The advantage of discriminative parameter estimation over generative parameter estimation is that discriminative training takes alternative (negative) examples into account as well as the correct (positive) examples. While generative training estimates a model that can generate the positive examples, discriminative training estimates model parameters that discriminate the positive examples from the negative ones. In ASR and MT tasks, positive examples are the correct transcriptions or translations and negative examples are the erroneous candidate transcriptions or translations. Discriminative models utilize these examples to optimize an objective function that is directly related to the system performance. Discriminative acoustic model training for ASR utilizes objective functions like Maximum Mutual Information (MMI) (Povey and Woodland 2000; Bahl et al. 1986) and Minimum Phone Error (MPE) (Povey and Woodland 2002) to estimate the acoustic model parameters that represent the discrimination between alternative classes. Discriminative language model (DLM) training for ASR aims to optimize the WER while learning the model parameters that discriminate the correct transcription of an utterance from the other candidate transcriptions. Another advantage of DLM is that discriminative language modeling is a feature-based approach, like conditional random fields (CRFs) (Lafferty et al. 2001) and maximum entropy models (Berger et al. 1996), therefore, it allows for easy integration of relevant knowledge sources, such as morphology, syntax, and semantics, into language modeling. As a result of improved parameter estimation with discriminative training and ease of incorporating overlapping features, discriminatively trained language models have been demonstrated to consistently outperform generative language modeling approaches (Roark et al. 2007, 2004; Collins et al. 2005; Shafran and Hall 2006).

In this section we will briefly explain the DLMs and the linguistically and statistically motivated features extracted at lexical and sub-lexical levels for Turkish DLMs.

4.6.1 Discriminative Language Model

This section describes the framework for discriminatively trained language models used for ASR. The definitions and notations given in Roark et al. (2007) are modified to match the definitions and notations of this chapter. The main components of DLMs are as follows:

1. **Training Examples:** These are the pairs (X_i, W_i) for $i = 1 \ldots N$. Here, X_i are the utterances and W_i are the corresponding reference transcriptions.

2. **GEN**(X): For each utterance X, this function enumerates a finite set of alternative hypotheses, represented as a lattice or N-best list output of the baseline ASR system of that utterance.
3. $\boldsymbol{\Phi(X, W)}$: A d-dimensional real-valued feature vector ($\Phi(X, W) \in \Re^d$). The representation $\boldsymbol{\Phi}$ defines the mapping from the (X, W) pair to the feature vector $\Phi(X, W)$. When the feature depends only on W, we simplify the notation to $\Phi(W)$ to match the notation used for other feature-based language models.
4. $\bar{\boldsymbol{\alpha}}$: A vector of discriminatively learned feature parameters ($\bar{\alpha} \in \Re^d$).

Like many other supervised learning approaches, DLM requires labeled input:output pairs as the training examples. Utterances with the reference transcriptions are utilized as the training examples, $(X_1, W_1) \ldots (X_N, W_N)$. These utterances are decoded with the baseline acoustic and language models in order to obtain the lattices or the N-best lists, in other words, the output of the $GEN(X)$ function. Since speech data with transcriptions are limited compared to the text data, it may not be possible to train the baseline acoustic and in-domain language models, and the DLM on separate corpora. Therefore, DLM training data is generated by breaking the acoustic training data into k-folds, and recognizing the utterances in each fold using the baseline acoustic model (trained on all of the utterances) and an n-gram language model trained on the other $k - 1$-folds to alleviate over-training of the language models. Acoustic model training is more expensive and less prone to over-training than n-gram language model training (Roark et al. 2007), so it is not typically controlled in the same manner.

Discriminative language modeling is a feature-based sequence modeling approach, where each element of the feature vector, $\Phi_0(X, W) \ldots \Phi_{d-1}(X, W)$, corresponds to a different feature. Each candidate hypothesis of an utterance has a score from the baseline acoustic and language models. This score is used as the first element of the feature vector, $\Phi_0(X, W)$. This feature is defined as the "log-probability of W in the lattice produced by the baseline recognizer for utterance X." In the scope of this chapter, the rest of the features depend only on W and will be denoted by $\Phi(W)$. The basic approach for the other DLM features is to use n-grams in defining features. The n-gram features are defined as the number of times a particular n-gram is seen in the candidate hypothesis. The details of the features used in Turkish DLMs will be explained in Sect. 4.6.2.

Each DLM feature has an associated parameter, i.e., α_i for $\Phi_i(X, W)$. The best hypothesis under the $\bar{\alpha}$ model, W^*, maximizes the inner product of the feature and the parameter vectors, as given in Eq. (4.10). The values of $\bar{\alpha}$ are learned in training and the best hypothesis under this model is searched for in decoding.

$$\begin{aligned} W^* &= \underset{W \in GEN(X)}{\operatorname{argmax}} \ \langle \bar{\alpha}, \Phi(X, W) \rangle \\ &= \underset{W \in GEN(X)}{\operatorname{argmax}} \ (\alpha_0 \Phi_0(X, W) + \alpha_1 \Phi_1(W) + \cdots + \alpha_{d-1} \Phi_{d-1}(W)) \end{aligned} \tag{4.10}$$

Inputs: Training examples (X_i, R_i) for $i = 1 \ldots N$
Initilization: $\bar{\alpha}_0^N = (\alpha_0, 0, \ldots, 0)$
Algorithm:
For $t = 1 \ldots T$
$\bar{\alpha}_t^0 = \bar{\alpha}_{t-1}^N$
For $i = 1 \ldots N$
$W^* = \underset{W \in GEN(X_i)}{\text{argmax}} \langle \bar{\alpha}_t^{i-1}, \bar{\Phi}(X_i, W) \rangle$
$\bar{\alpha}_t^i = \bar{\alpha}_t^{i-1} + \Phi(X_i, R_i) - \Phi(X_i, W^*)$
Output: Averaged parameters $\bar{\alpha} = \sum_{i,t} \bar{\alpha}_t^i / NT$

Fig. 4.4 A variant of the perceptron algorithm given in Roark et al. (2007). $\bar{\alpha}_t^i$ represents the feature parameters after the tth pass on the ith example. R_i is the gold-standard hypothesis

In basic DLM training, the parameters are estimated using a variant of the perceptron algorithm (shown in Fig. 4.4). The main idea in this algorithm is to penalize features associated with the current 1-best hypothesis, and to reward features associated with the gold-standard hypothesis (reference or lowest-WER hypothesis). It has been found that the perceptron model trained with the reference transcription as the gold-standard hypothesis is much more sensitive to the value of the α_0 constant (Roark et al. 2007). Therefore, we use the lowest-WER hypothesis (oracle) as the gold-standard hypothesis. Averaged parameters, $\bar{\alpha}_{AVG}$, are utilized in decoding held-out and test sets, since averaged parameters have been shown to outperform regular perceptron parameters in tagging tasks and also give much greater stability of the tagger (Collins 2002). See Roark et al. (2007) for the details of the training algorithm.

4.6.2 Feature Sets for Turkish DLM

This section describes the feature sets utilized in Turkish DLM experiments in the context of ASR (Arısoy et al. 2012; Sak et al. 2012). In order to generate the negative examples, we used a baseline Turkish ASR system to decode the DLM training set utterances yielding an N-best list for each training utterance. We then extracted the features from the correct transcriptions of the utterances together with the N-best list outputs of the baseline ASR system. In this section we investigate linguistically and statistically motivated features in addition to the basic n-gram features extracted from the word and sub-lexical ASR hypotheses.

4.6.2.1 Basic *n*-Gram Features

The basic n-gram features consist of word n-gram features extracted from word ASR hypotheses and sub-lexical n-gram features extracted from sub-lexical ASR hypotheses. Consider the Turkish phrase "`derneklerinin öncülüğünde`" given in Fig. 4.3. The unigram and bigram word features extracted from this phrase

are as follows:

$\Phi_i(W)$ = number of times "derneklerinin" is seen in W
$\Phi_j(W)$ = number of times "öncülüğünde" is seen in W
$\Phi_k(W)$ = number of times "derneklerinin öncülüğünde" is seen in W

We use a statistical morph-based ASR system to obtain the sub-lexical ASR hypotheses from which we extract the basic sub-lexical n-gram features. Some examples of the morph unigram and bigram features for the phrase in Fig. 4.3 are given as follows:

$\Phi_i(W)$ = number of times "dernek" is seen in W
$\Phi_j(W)$ = number of times "+lerinin" is seen in W
$\Phi_k(W)$ = number of times "öncü +lüğü" is seen in W

where the non-initial morphs were marked with "+" in order to find the word boundaries easily after recognition.

4.6.2.2 Linguistically Motivated Features

This section describes the morphological and syntactic features utilized in Turkish DLM. The rich morphological structure of Turkish introduces challenges for ASR systems (see Sect. 4.3). We aim to turn this challenging structure into a useful information source when reranking N-best word hypotheses with DLMs. Therefore, we utilize information extracted from morphological decompositions as DLM features. In our work, we have used root and stem+ending n-grams as the morphological features. In order to obtain the features, we first morphologically analyzed and disambiguated all the words in the hypothesis sentences using a morphological parser (Sak et al. 2011). The words that cannot be analyzed with the parser are left unparsed and represented as nominal nouns.

In order to obtain the root n-gram features, we first replace the words in the hypothesis sentences with their roots using the morphological decompositions. Then we generate the n-gram features as before, treating the roots as words. The root unigram and bigram features, with examples from Fig. 4.3, are listed below:

$\Phi_i(W)$ = number of times "dernek" is seen in W
$\Phi_j(W)$ = number of times "öncü" is seen in W
$\Phi_k(W)$ = number of times "dernek öncü" is seen in W

For the stem+ending n-gram features, we first extract the stem from the morphological decomposition and take the remaining part of the word as the ending. If there is no ending in the word, we use a special symbol to represent the empty ending. After converting the hypothesis sentences to stem and ending sequences, we generate the n-gram features in the same way with words as if stems and endings were words. The stem+ending unigram and bigram features, with examples from

Fig. 4.3, are listed below:

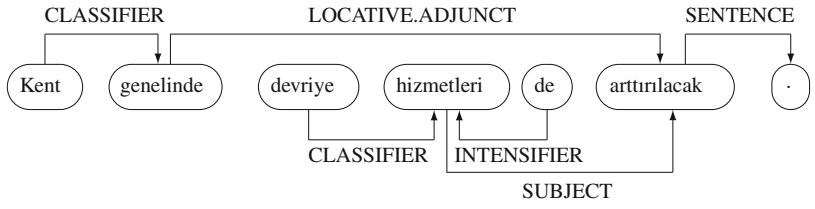
$\Phi_j(W)$ = number of times "+lerinin" is seen in W
$\Phi_k(W)$ = number of times "öncü +lüğünde" is seen in W

Syntax is an important information source for language modeling due to its role in sentence formation. Syntactic information has been incorporated into conventional generative language models using left-to-right parsers to capture long distance dependencies in addition to $n-1$ previous words (Chelba and Jelinek 2000; Roark 2001). Feature-based reranking approaches (Collins et al. 2005; Rosenfeld et al. 2001; Khudanpur and Wu 2000) also make use of syntactic information. The success of these approaches lead us to investigate syntactic features for Turkish DLMs.

For the syntactic DLM features, we explored feature definitions similar to Collins et al. (2005). We used part-of-speech tag n-grams and head-to-head (H2H) dependency relations between lexical items or their part-of-speech tags as the syntactic features. Part-of-speech tag features were utilized in an effort to obtain class-based generalizations that may capture well-formedness tendencies. H2H dependency relations were utilized since the presence of a word or morpheme can depend on the presence of another word or morpheme in the same sentence and this information is represented in the dependency relations.

The syntactic features will be explained with the dependency analysis given in Fig. 4.5 for a Turkish sentence, which translates as "Patrol services will also be increased throughout the city." The incoming and outgoing arrows in the figure show the dependency relations between the head and the dependent words with the type of the dependency. The words with English glosses, part-of-speech tags associated with the words are also given in the example. The dependency parser by Eryiğit et al. (2008) was used for the dependency analysis.

To obtain the syntactic features from the training examples, we first generated the dependency analyses of hypothesis sentences. Then we extracted the part-of-speech tag and H2H features from these dependency analyses. Here, it is important to note that hypothesis sentences contain recognition errors and the parser generates



WORD:	Kent	genelinde	devriye	hizmetleri	de	arttırılacak	.
GLOSS:	The city	throughout	patrol	services	also	will be increased	.
POS:	[Noun]	[Noun]	[Noun]	[Noun]	[Conj]	[Verb]	[Punc]

Fig. 4.5 Example dependency analysis for syntactic features

the best possible dependency relations even for incorrect hypotheses. The syntactic features are listed below with examples from Fig. 4.5.

- Part-of-speech tag n-gram features:
  Example for the word '`Kent`':
  $\Phi_k(W)$ = number of times "`[Noun]`" is seen in W
  Example for the words '`hizmetleri de`':
  $\Phi_k(W)$ = number of times "`[Noun] [Conj]`" is seen in W

- Head-to-Head (H2H) dependencies:
  Examples for the words '`Kent genelinde`':
  - dependencies between lexical items:
    $\Phi_k(W)$ = number of times "`CLASSIFIER Kent genelinde`" is seen in W
  - dependencies between a single lexical item and the part-of-speech of another item:
    $\Phi_k(W)$ = number of times "`CLASSIFIER Kent [Noun]`" is seen in W
    $\Phi_l(W)$ = number of times "`CLASSIFIER [Noun] genelinde`" is seen in W
  - dependencies between part-of-speech tags of lexical items:
    $\Phi_k(W)$ = number of times "`CLASSIFIER [Noun] [Noun]`" is seen in W

4.6.2.3 Statistically Motivated Features

The advantage of statistical sub-lexical units compared to their linguistic counterparts is that they do not require linguistic knowledge for word segmentation. As a result, statistical morphs do not convey explicit linguistic information like morphemes and obtaining linguistic information from morph sequences is not obvious. One way of information extraction from morphs is to convert them into word-like units and to apply the same procedure with words. However, this indirect approach tends to fail when concatenation of morph sequences does not generate morphologically correct words. In addition, this approach contradicts with the main idea of statistical morphs—obtaining sub-lexical units without any linguistic tools. Therefore, we focused on exploring representative features of implicit morpho-syntactic information in morph sequences. We explored morph-based features similar to part-of-speech tag and H2H dependency features using data driven approaches.

The first feature set is obtained by clustering morphs. We applied two hierarchical clustering approaches on morphs to obtain meaningful categories. The first one is the algorithm by Brown et al. (1992) which aims to cluster words into semantically-based or syntactically-based groupings by maximizing the average mutual information of adjacent classes. Brown et al.'s algorithm is proposed for

class-based n-gram language models and the optimization criterion in clustering is directly related to the n-gram language model quality. Utilizing n-gram features in DLMs makes this clustering an attractive approach for our investigation. The second approach utilizes minimum edit distance (MED) as the similarity function in bottom-up clustering. The motivation in this algorithm is to capture the syntactic similarity of morphs using their graphemic similarities, since a non-initial morph can cover a linguistic morpheme, a group of morphemes or pieces of morphemes. In our application, we modify MED to softly penalize the variations in the lexical and surface forms of morphemes. Note that this clustering is only meaningful for non-initial morphs since graphemic similarity of initial morphs does not reveal any linguistic information. Therefore, we only cluster the non-initial morphs and all the initial morphs are assigned to the same cluster with MED-based clustering approach.

Clustering is applied to morph sequences and each morph is assigned to one of the predetermined number of classes. The class associated with a particular morph is considered as the tag of that morph and utilized in defining DLM features. Clustering-based features are defined in a similar way with part-of-speech tag n-gram features, the class labels of morphs playing the role of the part-of-speech tags of words.

The second feature set is obtained with the triggering information obtained from morph sequences. These features are motivated by the H2H dependency features in words. Considering initial morphs as stems and non-initial morphs as suffixes, we assume that the existence of a morph can trigger another morph in the same sentence. The morphs in trigger pairs are believed to co-occur for a syntactic function, like the syntactic dependencies of words, and these pairs are utilized to define the long distance morph trigger features. Long distance morph trigger features are similar to the trigger features proposed in Rosenfeld (1994) and Singh-Miller and Collins (2007). We only consider sentence level trigger pairs to capture the syntactic-level dependencies instead of discourse-level information. The candidate morph trigger pairs are extracted from the hypothesis sentences (1-best and oracle) to obtain also the negative examples for DLMs. An example morph hypothesis sentence with the candidate trigger pairs is given in Fig. 4.6. Among the possible candidates, we try to select only the pairs where morphs are occurring together for a special function. This is formulated with hypothesis testing where the null hypothesis (H_0) represents the independence and the alternative hypothesis

Morph hypothesis:
dernek +lerinin öncü +lüğü +nde
Candidate trigger pairs:
dernek +lerinin dernek öncü dernek +lüğü dernek +nde
+lerinin öncü +lerinin +lüğü +lerinin +nde
öncü +lüğü öncü +nde
+lüğü +nde

Fig. 4.6 A morph hypothesis sentence and the candidate trigger pairs extracted from this hypothesis

(H_1) represents the dependence assumptions of morphs in the pairs (Manning and Schütze 1999). The pairs with higher likelihood ratios ($\log \frac{L(H_1)}{L(H_0)}$) are assumed to be the morph triggers and utilized as features. The number of times a morph trigger pair is seen in the candidate hypothesis is defined as long-distance trigger features. For instance, if among the candidate trigger pairs, given in Fig. 4.6, "öncü +lüğü" is selected as a trigger pair, the feature for this pair is defined as follows:

$\Phi_k(W)$ = number of times "öncü +lüğü" is seen in W

4.7 Conclusions

In this chapter, we summarized the language modeling research for Turkish text and speech processing applications. The agglutinative nature of Turkish results in high OOV rates which can be alleviated by using sub-lexical units for language modeling. Knowledge-based linguistic methods and data-driven unsupervised statistical methods have both been used for segmenting words into sub-lexical units. Language models based on these units have advantages of those based on words and often result in improved performance. After many years of research, n-gram language models are still the most popular language modeling technique. However, in certain applications such as ASR, discriminative language models have been shown to improve the task performance. The ASR performance of the language models presented in this chapter is provided in Chap. 5. Despite significant progress in the recent years, language modeling for morphologically rich languages such as Turkish remains an active field of research.

References

Alumäe T (2006) Methods for Estonian large vocabulary speech recognition. PhD thesis, Tallinn University of Technology, Tallinn

Arısoy E (2004) Turkish dictation system for radiology and broadcast news applications. Master's thesis, Boğaziçi University, Istanbul

Arısoy E, Saraçlar M (2009) Lattice extension and vocabulary adaptation for Turkish LVCSR. IEEE Trans Audio Speech Lang Process 17(1):163–173

Arısoy E, Dutağacı H, Arslan LM (2006) A unified language model for large vocabulary continuous speech recognition of Turkish. Signal Process 86(10):2844–2862

Arısoy E, Sak H, Saraçlar M (2007) Language modeling for automatic Turkish broadcast news transcription. In: Proceedings of INTERSPEECH, Antwerp, pp 2381–2384

Arısoy E, Can D, Parlak S, Sak H, Saraçlar M (2009a) Turkish broadcast news transcription and retrieval. IEEE Trans Audio Speech Lang Process 17(5):874–883

Arısoy E, Pellegrini T, Saraçlar M, Lamel L (2009b) Enhanced Morfessor algorithm with phonetic features: application to Turkish. In: Proceedings of SPECOM, St. Petersburg

Arısoy E, Saraçlar M, Roark B, Shafran I (2012) Discriminative language modeling with linguistic and statistically derived features. IEEE Trans Audio Speech Lang Process 20(2):540–550

Bahl L, Brown P, deSouza P, Mercer R (1986) Maximum mutual information estimation of Hidden Markov Model parameters for speech recognition. In: Proceedings of ICASSP, Tokyo, pp 49–52

Bayer AO, Çiloğlu T, Yöndem MT (2006) Investigation of different language models for Turkish speech recognition. In: Proceedings of IEEE signal processing and communications applications conference, Antalya, pp 1–4

Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155

Berger AL, Della Pietra SD, Della Pietra VJD (1996) A maximum entropy approach to natural language processing. Comput Linguist 22(1):39–71

Bilmes JA, Kirchhoff K (2003) Factored language models and generalized parallel backoff. In: Proceedings of NAACL-HLT, Edmonton, pp 4–6

Brent MR (1999) An efficient, probabilistically sound algorithm for segmentation and word discovery. Mach Learn 34:71–105

Brown PF, Pietra VJD, deSouza PV, Lai JC, Mercer RL (1992) Class-based *n*-gram models of natural language. Comput Linguist 18(4):467–479

Çarkı K, Geutner P, Schultz T (2000) Turkish LVCSR: towards better speech recognition for agglutinative languages. In: Proceedings of ICASSP, Istanbul, pp 1563–1566

Chelba C, Jelinek F (2000) Structured language modeling. Comput Speech Lang 14(4):283–332

Chen SF, Goodman J (1999) An empirical study of smoothing techniques for language modeling. Comput Speech Lang 13(4):359–394

Choueiter G, Povey D, Chen SF, Zweig G (2006) Morpheme-based language modeling for Arabic. In: Proceedings of ICASSP, Toulouse, pp 1052–1056

Çiloğlu T, Çömez M, Şahin S (2004) Language modeling for Turkish as an agglutinative language. In: Proceedings of IEEE signal processing and communications applications conference, Kuşadası, pp 461–462

Collins M (2002) Discriminative training methods for Hidden Markov Models: theory and experiments with perceptron algorithms. In: Proceedings of EMNLP, Philadelphia, PA, pp 1–8

Collins M, Saraçlar M, Roark B (2005) Discriminative syntactic language modeling for speech recognition. In: Proceedings of ACL, Ann Arbor, MI, pp 507–514

Creutz M, Lagus K (2002) Unsupervised discovery of morphemes. In: Proceedings of the workshop on morphological and phonological learning, Philadelphia, PA, pp 21–30

Creutz M, Lagus K (2005) Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Publications in Computer and Information Science Report A81, Helsinki University of Technology, Helsinki

Dutağacı H (2002) Statistical language models for large vocabulary continuous speech recognition of Turkish. Master's thesis, Boğaziçi University, Istanbul

Erdoğan H, Büyük O, Oflazer K (2005) Incorporating language constraints in sub-word based speech recognition. In: Proceedings of ASRU, San Juan, PR, pp 98–103

Eryiğit G, Nivre J, Oflazer K (2008) Dependency parsing of Turkish. Comput Linguist 34(3):357–389

Goldsmith J (2001) Unsupervised learning of the morphology of a natural language. Comput Linguist 27(2):153–198

Hacıoğlu K, Pellom B, Çiloğlu T, Öztürk Ö, Kurimo M, Creutz M (2003) On lexicon creation for Turkish LVCSR. In: Proceedings of EUROSPEECH, Geneva, pp 1165–1168

Hakkani-Tür DZ (2000) Statistical language modeling for agglutinative languages. PhD thesis, Bilkent University, Ankara

Harris Z (1967) Morpheme boundaries within words: report on a computer test. In: Transformations and discourse analysis papers, vol 73. University of Pennsylvania, Philadelphia, PA

Hetherington IL (1995) A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA

Hirsimäki T (2009) Advances in unlimited vocabulary speech recognition for morphologically rich languages. PhD thesis, Helsinki University of Technology, Espoo

Hirsimäki T, Creutz M, Siivola V, Kurimo M, Virpioja S, Pylkkönen J (2006) Unlimited vocabulary speech recognition with morph language models applied to Finnish. Comput Speech Lang 20(4):515–541

Jelinek F (1997) Statistical methods for speech recognition. The MIT Press, Cambridge, MA

Kanevsky D, Roukos S, Sedivy J (1998) Statistical language model for inflected languages. US patent No: 5,835,888

Khudanpur S, Wu J (2000) Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling. Comput Speech Lang 14:355–372

Kirchhoff K, Vergyri D, Bilmes J, Duh K, Stolcke A (2006) Morphology-based language modeling for conversational Arabic speech recognition. Comput Speech Lang 20(4):589–608

Kneissler J, Klakow D (2001) Speech recognition for huge vocabularies by using optimized sub-word units. In: Proceedings of INTERSPEECH, Aalborg, pp 69–72

Kurimo M, Puurula A, Arısoy E, Siivola V, Hirsimäki T, Pylkkönen J, Alumäe T, Saraçlar M (2006) Unlimited vocabulary speech recognition for agglutinative languages. In: Proceedings of NAACL-HLT, New York, NY, pp 487–494

Kwon OW, Park J (2003) Korean large vocabulary continuous speech recognition with morpheme-based recognition units. Speech Comm 39:287–300

Lafferty JD, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML, Williams, MA, pp 282–289

Manning C, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, MA

Mengüşoğlu E, Deroo O (2001) Turkish LVCSR: database preparation and language modeling for an agglutinative language. In: Proceedings of ICASSP, Salt Lake City, UT, pp 4018–4021

Mihajlik P, Fegyò T, Tüske Z, Ircing P (2007) A morpho-graphemic approach for the recognition of spontaneous speech in agglutinative languages like Hungarian. In: Proceedings of INTER-SPEECH, Antwerp, pp 1497–1500

Pellegrini T, Lamel L (2007) Using phonetic features in unsupervised word decompounding for ASR with application to a less-represented language. In: Proceedings of INTERSPEECH, Antwerp, pp 1797–1800

Pellegrini T, Lamel L (2009) Automatic word decompounding for ASR in a morphologically rich language: application to amharic. IEEE Trans Audio Speech Lang Process 17(5):863–873

Podvesky P, Machek P (2005) Speech recognition of Czech – inclusion of rare words helps. In: Proceedings of the ACL student research workshop, Ann Arbor, MI, pp 121–126

Povey D, Woodland PC (2000) Large-scale MMIE training for conversational telephone speech recognition. In: Proceedings of NIST speech transcription workshop, College Park, MD

Povey D, Woodland PC (2002) Minimum phone error and i-smoothing for improved discriminative training. In: Proceedings of ICASSP, Orlando, FL, pp 105–108

Roark B (2001) Probabilistic top-down parsing and language modeling. Comput Linguist 27(2):249–276

Roark B, Saraçlar M, Collins MJ, Johnson M (2004) Discriminative language modeling with conditional random fields and the perceptron algorithm. In: Proceedings of ACL, Barcelona, pp 47–54

Roark B, Saraçlar M, Collins M (2007) Discriminative *n*-gram language modeling. Comput Speech Lang 21(2):373–392

Rosenfeld R (1994) Adaptive statistical language modeling: a maximum entropy approach. PhD thesis, Carnegie Mellon University, Pittsburgh, PA

Rosenfeld R (1995) Optimizing lexical and *n*-gram coverage via judicious use of linguistic data. In: Proceedings of EUROSPEECH, Madrid, pp 1763–1766

Rosenfeld R, Chen SF, Zhu X (2001) Whole-sentence exponential language models: a vehicle for linguistic-statistical integration. Comput Speech Lang 15(1):55–73

Rotovnik T, Maučec MS, Kačic Z (2007) Large vocabulary continuous speech recognition of an inflected language using stems and endings. Speech Commun 49(6):437–452

Sak H, Güngör T, Saraçlar M (2007) Morphological disambiguation of Turkish text with perceptron algorithm. In: Proceedings of CICLING, Mexico City, pp 107–118

Sak H, Saraçlar M, Güngör T (2010) Morphology-based and sub-word language modeling for Turkish speech recognition. In: Proceedings of ICASSP, Dallas, TX, pp 5402–5405

Sak H, Güngör T, Saraçlar M (2011) Resources for Turkish morphological processing. Lang Resour Eval 45(2):249–261

Sak H, Saraçlar M, Güngör T (2012) Morpholexical and discriminative language models for Turkish automatic speech recognition. IEEE Trans Audio Speech Lang Process 20(8):2341–2351

Schwenk H (2007) Continuous space language models. Comput Speech Lang 21(3):492–518

Shafran I, Hall K (2006) Corrective models for speech recognition of inflected languages. In: Proceedings of EMNLP, Sydney, pp 390–398

Siivola V, Hirsimäki T Teemu, Creutz M, Kurimo M (2003) Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In: Proceedings of EUROSPEECH, Geneva, pp 2293–2296

Singh-Miller N, Collins M (2007) Trigger-based language modeling using a loss-sensitive perceptron algorithm. In: Proceedings of ICASSP, Honolulu, HI, pp 25–28

Wang W, Harper MP (2002) The SuperARV language model: investigating the effectiveness of tightly integrating multiple knowledge sources. In: Proceedings of EMNLP, Philadelphia, PA, pp 238–247

Whittaker E, Woodland P (2000) Particle-based language modelling. In: Proceedings of ICSLP, Beijing, vol 1, pp 170–173