Half-Context Language Models

Hinrich Schütze*
University of Stuttgart

Michael Walsh**
University of Stuttgart

This article investigates the effects of different degrees of contextual granularity on language model performance. It presents a new language model that combines clustering and half-contextualization, a novel representation of contexts. Half-contextualization is based on the half-context hypothesis that states that the distributional characteristics of a word or bigram are best represented by treating its context distribution to the left and right separately and that only directionally relevant distributional information should be used. Clustering is achieved using a new clustering algorithm for class-based language models that compares favorably to the exchange algorithm. When interpolated with a Kneser-Ney model, half-context models are shown to have better perplexity than commonly used interpolated n-gram models and traditional class-based approaches. A novel, fine-grained, context-specific analysis highlights those contexts in which the model performs well and those which are better treated by existing non-class-based models.

1. Introduction

Stochastic language models are a crucial component of many speech and language technology applications. The key problem encountered by these models is that sparse data make the accurate estimation of the probability of novel and rare word sequences difficult.

In order to address this, language model researchers have developed a number of strategies. Of particular interest in this article are the following four:

1. **Context length.** Careful selection of the length of the history or context that is the basis for predicting the next word.
2. **Interpolation.** Models typically interpolate several predictions, for example, predictions that are based on several different context lengths.
3. **Classes.** In a class-based model, prediction is (partially) based on classes that the *n*-grams involved are members of.
4. **Similarity.** Similarity models smooth predictions with predictions for similar entities.

* Institute for Natural Language Processing. E-mail: hinrichcl11@ifnlp.org.
** Universität Stuttgart, Institut für Maschinelle Sprachverarbei-Azenbergstrasse 12, D-70174 Stuttgart, Germany. E-mail: michael.walsh@ims.uni-stuttgart.de.

Submission received: 3 August 2010; revised submission received: 17 February 2011; accepted for publication: 30 March 2011

The language models that are most commonly used today, in particular the modified Kneser-Ney (KN) model (Chen and Goodman 1998), are based on the first two strategies, context length and interpolation—that is, they interpolate distributions of different history lengths. We will call such models **history-length interpolated models**. The set of contexts that history-length-interpolated models base their prediction on is limited to those whose history is *identical* for the history length considered. For example, the length-2 component of the model will compute the probability of $P(w_3|w_1w_2)$ based on contexts in the training corpus with identical history w_1w_2.

Class-based and similarity-based models consider a wider range of contexts. Their estimates rely on contexts in the training data that are similar to (or in the same class as) the new sequence whose probability is to be estimated. Thus, for example, in attempting to estimate a probability for the bigram *black cloud*, unseen in training, the transition probability associated with the class to which *black* belongs being followed by the class to which *cloud* belongs can be used. The intuition is that although *black cloud* might not have been seen in training, the class sequence containing related bigrams like *gray cloud*, or *black mist*, or *gray mist*, that is, combinations of other members of the two classes seen in training, can offer a reasonable estimate. In principle, this type of generalization is more powerful than history-length interpolation and has been, and continues to be, used to good effect in a variety of domains. However, the model must be a good model of the distribution of sequences of strings; if its assumptions are too unrealistic or approximate, then class-based generalization will perform worse than history-length interpolation.

Although there has been much work on class-based and similarity-based language models in recent years, no such model has been widely adopted as superior to history-length-interpolated models. We believe one reason for this is that the granularity of context that is optimal for generalization has not been investigated sufficiently. Consequently, in this article, we present the following contributions:

- We demonstrate that class-based models can be made more effective. In particular, we put forward the **half-context hypothesis** as a general principle on the basis of which to construct class-based language models.
- We argue for the novel, beneficial use of a mixed n-gram class of both bigrams and unigrams instead of a class of unigrams alone.
- We specify a discounting method which facilitates better treatment of rare events.
- We deploy a new clustering algorithm for class-based language models that is more efficient than the exchange algorithm.
- We perform a systematic investigation, including significance testing, of half-context versus whole-context class-based models which demonstrates the utility of a half-context approach.
- We carry out a novel fine-grained context-specific experimental validation of a half-context model that performs better than a traditional class-based model, and, when interpolated, improves on a modified KN trigram model. This new fine-grained analysis distinguishes those contexts best suited to history-length interpolation and those most appropriate for class-based generalization.

These contributions, particularly our analyses, offer a richer understanding of the relative characteristics of history-length interpolation and class-based generalization and should lead to more powerful language models that combine class-based and history-length generalization mechanisms.

The remainder of the article is organized as follows. Section 2 defines half-context representation and puts forward the half-context hypothesis. Section 3 develops a half-context language model in the context of a specific subset of related prior work on language modeling. Additional related work is discussed in a subsequent subsection. Parameter estimation is described in Section 4. A variety of models and interpolations are evaluated, and fine-grained results, significance tests, and context-specific analyses are discussed in Section 5. Conclusions and opportunities for future work are presented in Section 6.

2. Half Contexts

The representation employed in this article builds on a specification used in our earlier work (Schütze 1993, 1995; Schütze and Walsh 2008), motivated by Exemplar Theory (Hintzman 1986; Nosofsky 1986; Pierrehumbert 2001), where rich exemplar representations facilitated the acquisition of local grammatical knowledge and outperformed a categorical representation in the same task. Specifically, each word was represented in terms of its immediate left and right neighborhood context. These neighborhoods were treated separately for two reasons: (1) separate treatment of left-neighbor information and right-neighbor information resulted in reduced complexity in the model and better generalization, and (2) right and left context behavior can differ considerably, for example, *him* and *her* would have very similar left contexts but could have significantly differing right contexts (e.g., compare *the life in her garden* vs. *the life in him garden*).

These representations of left and right context distributions of a given word were known as **half-words** but can in fact be viewed as a word-level instantiation of a broader representational formalism which we term **half-contextualization**. According to this schema a given unit (word, *n*-gram, class, etc.) is represented in terms of **half-context (HC) distributions** over its immediate left and right neighborhoods. Hence, for example, at the bigram level, each bigram type is specified by two distributions, namely a left HC distribution P^l and right HC distribution P^r that capture the bigram's behavior to the immediate left/right. For example, given *walk home early* twice, and *drive home early* once, then the left HC distribution of the bigram *home early*, denoted $P^l_{home\ early}$, is $P^l_{home\ early}(walk) = 2/3$ and $P^l_{home\ early}(drive) = 1/3$, and 0 for all other words. These HC distributions underpin the HC language models presented in Section 3.

In order to determine the extent of the particular merits of considering words as possessing two separate directional behaviors, in the experiments that follow we compare our HC language model against a whole-context (WC) model where a given word's WC distribution is a single distribution which combines the word's left and right HC distributions. For a clear statement of the contrast HC vs. WC, we define inward and outward distributions. For the estimation of $P(w_{n+1}|w_{1,n})$ based on a training set S, the **inward distributions** $\mathrm{IW}_{w_{n+1}|w_{1,n}}$ consist of the set of *right* contexts of $w_{1,n}$ and *left* contexts of w_{n+1} in S; the **outward distributions** $\mathrm{OW}_{w_{n+1}|w_{1,n}}$ consist of the set of *left* contexts of $w_{1,n}$ and *right* contexts of w_{n+1} in S.

We can then state our underlying hypothesis that HC-based classes are better for language modeling than WC-based classes as follows.

> **Half-Context Hypothesis.** A distributional language model should base its estimate of $P(w_{n+1}|w_{1,n})$ on contexts $v_{1,n}v_{n+1}$ whose inward distributions $\mathrm{IW}_{v_{n+1}|v_{1,n}}$ are similar to $\mathrm{IW}_{w_{n+1}|w_{1,n}}$. Similarity of the outward distributions $\mathrm{OW}_{w_{n+1}|w_{1,n}}$ and $\mathrm{OW}_{v_{n+1}|v_{1,n}}$ should not be employed as a criterion for using or not using training set contexts $v_{1,n}v_{n+1}$ for the estimation of $P(w_{n+1}|w_{1,n})$.

An example for the intuition behind the HC hypothesis is the *him/her* example given earlier. When estimating $P(him|Mary\ helped)$, a context like *Mary helped her* should also be considered as evidence because even though the right contexts of *him* and *her* in the corpus are dissimilar, their left contexts are similar. The HC hypothesis states that we should only worry about similarity of the "relevant side" of the n-grams involved, that is we should only consider *inward* distributional information. Most clustering algorithms used for class-based language models, notably the exchange algorithm (Brown et al. 1992; Kneser and Ney 1993; Martin, Liermann, and Ney 1998), are "whole-context" clustering algorithms that violate the hypothesis.

The HC hypothesis provides an alternative basis for designing class-based language models. In general, in designing a language model only information sources that are relevant for the task to be solved should be included. Adding additional complexity or nonrelevant additional features increases the variance of predictions without improving their accuracy. We can view this as a type of bias–variance tradeoff. Half-context models are simpler and have less variance because they only use one half of the available context information, the half that is actually useful for prediction. The experimental results that we report later in this article confirm this by demonstrating that half-context models perform significantly better than whole-context models.

A consequence of only using inward distributions in accordance with the half-context hypothesis is that we need two different types of classes: one set of classes for the predictors and another set of classes for the predictees. The reason is that we use two distinct and unrelated representations, left-context distributions to induce classes of predictees and right-context distributions to induce classes of predictors. In other words, half-context models are inherently asymmetric, reflecting the fact that language models are inherently asymmetric: The role of the predictor and the predicted are different. This asymmetry shows up in word-based models to a limited extent: In most models the unit of prediction is a word; predictors include n-grams of any size in principle, not just words. However, in a class-based model the asymmetry between predictor and predicted is more important: There is no justification for the premise (made, for example, in the Brown model) that the classes that are optimal for predictors are also the classes that are optimal for predictees. This observation has also been made by Gao et al. (2002), albeit without explicit reference to half contexts. We view our approach as better motivated since the asymmetry of our model is not posited, but follows from an analysis of the information sources needed for probabilistic inference in language modeling.

Linguistic theory also provides evidence for half-context models. In many theories, there is a single formal concept that can be instantiated either by arguments of prepositions or by arguments of transitive verbs. For example, there are few if any syntactic differences between the arguments that can appear after a preposition like *by* and after a transitive verb like *brought*. Thus, the predicting histories *by* and *brought* should be treated alike in a class-based model. But that is not possible in a

whole-context model. We can interpret this as a syntactic justification for half-context models. Referring back to our earlier allusion to the bias–variance tradeoff, if we had unlimited data, then estimating separate distributions for *by* and *brought* would be unproblematic; but because training sets are not unlimited, we can improve generalization by assigning the two linguistically identical contexts to the same right-context class.

Finally, efficiency is also a strong argument for half-context models. Time of clustering and storage requirements are cut in half by omitting those parts of the context that are nonrelevant. The time complexity of many clustering algorithms depends on the number of different types of features that occur in a particular cluster as opposed to the number of tokens. The number of feature types occurring in a cluster is reduced substantially in half-context models.

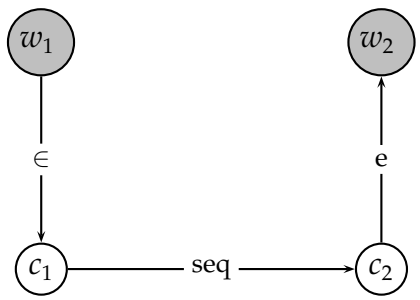
It is of course possible to find cases where the outward distributions are helpful for accurate estimation. Consider estimating $P(is|strilp)$, where *strilp* is a word that occurred once in the training set. Suppose for the sake of argument that *is* after nouns is more likely than *is* after adjectives (because phrases like *yellow is the new black* are infrequent). If *strilp* occurred in the context *a very strilp car*, then it is likely to be an adjective and $P(is|strilp)$ is low. If *strilp* occurred in the context *the strilp car*, then it could also be a noun (as in *the bakery car* or *the wedding ring*) and $P(is|strilp)$ should be estimated to be higher. In this case, it is the outward distribution of *strilp* that helps us to arrive at an accurate estimate. However, our hypothesis is not that there are no such cases; rather, we believe that as a generalization mechanism, only inward distributional information is useful in improving performance. This is borne out by the experiments reported herein.

In the future, there may be non-distributional models that use more complex inferences for language modeling. Parsing-based language models (e.g., Hall and Johnson 2003) are a first step in this direction. The hypothesis would probably not apply to such non-distributional models.

3. Half-Context Language Model

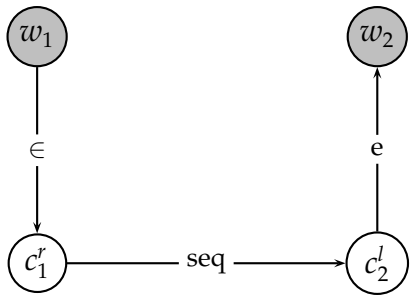
3.1 Half-Contextualization

Our starting point is the model by Brown et al. (1992). It models the probability of class c_2 following class c_1 where c_2 emits (e in the diagram) the member word w_2 and w_1 belongs to ($\in$, a deterministic process) class c_1:



This model has been frequently investigated and discussed. Recent examples include its successful application in word co-occurrence and sentence retrieval investigations (Momtazi and Klakow 2009; Momtazi, Khudanpur, and Klakow 2010), and polarity classification of movie reviews (Wiegand and Klakow 2008).

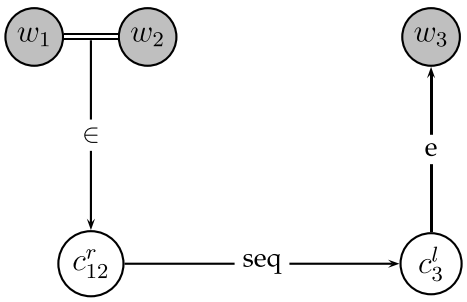
The key concept introduced in this article is that of a half-context class. Class-based language models can be *half-contextualized* by replacing classes that model right and left contexts simultaneously by right half-context classes c_1^r and left half-context classes c_2^l. Words are assigned to HC classes and these HC classes then generate words:



3.2 Mixed *n*-gram Classes

A second modification of the Brown model we propose is motivated by the fact that trigram models perform better than bigram models because a sequence of two words significantly limits the possible ways of continuing. For this reason, we condition the sequential continuation on a *mixed* n-*gram class* of both bigrams and unigrams instead of on a class of unigrams alone. The resulting model, the HC model, is depicted in Figure 1. We show the bigram w_1w_2 as the member of the class c_{12}^r, but c_{12}^r can also be the class of w_2 if w_1w_2 was not frequent enough to be included in the clustering (criteria for inclusion are discussed in Section 4).

To summarize, the generative process shown in Figure 1 is that the right-context class c_{12}^r to which the bigram w_1w_2 belongs generates a unigram left-context class c_3^l which generates w_3. As will become apparent from the description of parameter estimation and the clustering algorithm in Section 4, the HC classes in the model are based



$$P_{\text{HC}}(w_3|w_1w_2) = P_e(w_3|c_3^l)P_{\text{seq}}(c_3^l|c_{12}^r(w_1w_2))$$

Figure 1
The half-context language model.

on inward (IW) distributions only. The corresponding outward (OW) distributions are not taken into account in accordance with the HC hypothesis.

In addition to the novel use of half-contexts it is important to note that the right HC classes employed are mixed classes of unigrams and bigrams rather than of unigrams only. To our knowledge this represents the first such usage and can be motivated by the fact that frequent bigrams in language often behave similarly whereas the constituent unigrams do not. For example, the right HCs of the bigrams *University of* and *based in* are similar because both are often followed by locations; but the right HCs of *of* and *in* are much more diffuse and there are many prepositional objects that occur more often with *of* than with *in* (e.g., names of people) and others that occur more often with *in* than with *of* (e.g., *response*).

3.3 Discounting

In initial experiments, we found that it was difficult to achieve an improvement using class-based generalization because for many contexts history-length interpolation is the better strategy for estimation. For a high-frequency event, it can be best to base estimates on instances of this event with identical history only—instead of smoothing them with other contexts that are in the same class.

Consider the unigram *Hong*. In 3,998 out of 4,045 cases in the training set part of our corpus of *Wall Street Journal* (WSJ) articles (consisting of 40 million words), it is followed by *Kong*. In this case, redistributing probability mass to other members of the class that *Kong* is a member of will decrease the estimate for $P(Kong|Hong)$ (an estimate that should be close to 3,998/4,045) and decrease the model's performance. On the other hand, we have $H(P(w|Mr.)) \approx 11.9$ in our WSJ training set. Any of a large number of first and last names can occur after *Mr.* and a language model should reallocate some probability mass from names that did occur in this environment in the training set to those names that did not.

To treat these two different cases correctly, we use a variant of absolute discounting (Ney, Essen, and Kneser 1994). Following the notation of Chen and Goodman (1998), we first define the number $N_{1+}(w_{1,n}\bullet)$ of distinct words that can occur after an n-gram in the training set:

$$N_{1+}(w_{1,n}\bullet) = |\{w|C(w_{1,n}w) > 0\}|$$

where $C(w_{1,n})$ is the frequency of $w_{1,n}$ in the training set.

We then define the exemplar-theoretic (ET) language model as follows:

$$\begin{aligned} P_{\mathrm{ET}}(w_{n+1}|w_{1,n}) = {} & D\frac{N_{1+}(w_{1,n}\bullet)}{C(w_{1,n})}P_{\mathrm{HC}}(w_{n+1}|w_{1,n}) \\ & + \frac{\max(0, C(w_{1,n}w_{n+1}) - D)}{C(w_{1,n})} \end{aligned} \tag{1}$$

The discount D is a parameter of the model that controls how much of each count is redistributed to the class-based model. In a way that is similar to other discounting methods, this formalization satisfies the two desiderata stated earlier: The estimates of high-frequency events are, in relative terms, much less affected than those of low-frequency events.

P_{ET} is the exemplar-theoretic model we will evaluate in the experiments described below. We use an analogous model for WC distributions. In that case, P_{HC} is simply replaced by P_{WC} in Equation (1).

To summarize, the innovations of our exemplar-theoretic model are (1) the use of HC classes instead of WC classes, (2) the use of mixed classes of unigrams and bigrams (instead of classes of unigrams), and (3) the use of absolute discounting to concentrate the effect of class-based generalization on rare hard-to-estimate events while leaving robust estimates based on frequent events largely unchanged.

3.4 Additional Related Work

In addition to the motivating articles discussed earlier, other relevant work includes the randomization techniques applied by Emami and Jelinek (2005) to class-based *n*-gram language models. Half-context clusters are not at odds with a randomized approach as they could easily be implemented in such a fashion.

Other related research includes the "mixed" model employed by Uszkoreit and Brants (2008), in which a word bigram (as opposed to a class of bigrams) probabilistically generates a class. They use, in our terminology, whole-context classes. The experiments reported subsequently suggest that HC classes are preferable to WC classes in the Brown-type set-up (classes generating classes); we plan to investigate whether this is also true in a mixed model in future work.

Bassiou and Kotropoulos (2011) investigate two word-clustering techniques that operate on long-distance bigram probabilities (of varying distances) within a context and on interpolated long-distance bigram probabilities, both with a view to capturing long-distance dependencies. Evaluation of both clustering techniques—hierarchical clustering exploiting Mahalanobis distances to form compact clusters and Probabilistic Latent Semantic Analysis—demonstrates that the use of long distance bigrams or their interpolated varieties yield more compact and meaningful (in the case of interpolated long distance bigrams) word clusters than the use of the traditional bigram (and bigrams which employ trigger pairs over various histories). This research demonstrates an interesting avenue for contemporary models of word clustering and it would be no doubt interesting to see how such clustering strategies might contribute to half-context clustering, how their clusters would compare to those produced via bisecting *k*-means (though we cluster bigrams also), as proffered here, and indeed how long distance bigrams could be half-contextualized; these questions, however, are beyond the scope of the current article which seeks primarily to investigate the potential merits of half-contextualization.

Related work by Justo and Torres (2009) explores the use of language models that employ classes containing phrases. They describe their models as *two-level* because specific language models act within the classes. Their first approach takes into account the probabilities between words which constitute the different phrases of a given class, that is, phrases are sequences of unconnected words and words are considered the basic lexical unit, and their second approach considers phrases to be indivisible lexical units. The first model is also interpolated with a standard word-based language model, and the second model is interpolated with a standard phrase-based model. Word-error-rate analyses in an ASR system indicate that these models are better than their traditional counterparts. These results provide useful motivation for extending class-based language models from classes of isolated words to classes of longer sequences, such as the classes of bigrams employed in the half-context model. Their research

does not, however, consider the different directional behaviors of words or bigrams as we do.

Zitouni and Zhou (2007, 2008) propose linearly interpolated hierarchical language models (and a back-off variety [Zitouni 2007]) where each vocabulary item constitutes a leaf node in a word-tree, words are clustered into classes, and, in a recursive process, classes are clustered into more general classes until the root is reached. The tree root is a class containing all vocabulary items. In attempting to estimate the likelihood of an n-gram event they linearly interpolate over different language models, each one of which is trained on one level of the tree. In this way they seek to strike a balance between specificity and generalization. In constructing the class hierarchy, words are represented by their probability given their left and right neighboring words over a vocabulary (equivalent to the whole-context representation discussed in this article) and similarity between words is established using the Kullback-Leibler distortion measure. Words occurring frequently in similar contexts should be clustered together with a view to finding a set of clusters that minimizes global discriminative information (see also Bai et al. 1998). The clustering algorithm is based on k-means. The use of a hierarchical tree, and interpolating over it, represents an interesting approach not at odds with our research (i.e., that is, half-context classes could form nodes in the tree), although our approach differs in the separate treatment of word contexts, the use of bigrams as class members, and in the clustering methodology.

Additional related work includes research by Bahrani et al. (2008), who build class-based models using the k-means algorithm and words represented in terms of vectors where each vector element corresponds to the number of times the word had a particular part of speech tag given a tagged corpus. This approach would typically yield much shorter feature vectors than approaches (including our own) which have vectors matching the vocabulary size, thus leading to lower time complexity. They do not, however, avail of classes of bigrams as we do, nor look at directional behavior of words (though half-contextualization using part of speech tags would be an interesting extension of both our model and theirs). Abdoos and Naeini (2008) use a clustering ensemble approach to categorize words, although it is unclear from their evaluation how such an approach compares, in terms of performance, to others in the literature. Gao et al. (2002) propose an asymmetric clustering model (ACM) grounded upon the apt observation that different clusters for predicted and conditional words should be employed, a view shared here. Their research does not present an explicit treatment of half-contextualization, however, nor considers half-contextualization and the significance of inward distributional information as insights which meet language modeling needs. Furthermore, our evaluation also differs in that it involves comparison against a whole-context model and a modified KN trigram model, rather than a simple word trigram model. Our use of a mixed n-gram class of both bigrams and unigrams also represents a marked difference between approaches.

With regard to context direction Essen and Steinbiss (1992) also look at left and right contexts similarly to our approach. However, they do not compare half-context with whole-context approaches and they pursue a less efficient similarity-based approach in contrast to the class-based approach proposed here.

Finally, Dagan, Lee, and Pereira's (1999) similarity-based language model uses a similar word to the observed word as the conditioning context used to generate the next word in the sequence. Again, no comparison to whole-context approaches is made. A similarity-based approach is also difficult to use for large corpora as it would necessitate the calculation of similarity of every word to every other word in the corpus. Similarities

can be computed more efficiently for a subset of words on a smaller corpus, but then many of the rare events that class and similarity based methods are most beneficial for will not be covered. Our analyses in Section 5.2 and Section 5.3 demonstrate that half-context modeling is most beneficial for rare events. Similar concerns apply to other similarity-based models, such as those proposed by Bengio et al. (2003) and Schwenk and Koehn (2008).

4. Parameter Estimation

In this section we describe how the parameters of the model in Figure 1 are estimated. These parameters belong to two broad categories, namely, those which model the HC distributions (P^r and P^l) and are used in the construction of clusters, and those which capture emission probabilities P_e and sequence probabilities P_{seq} that are used when the model is applied. Estimates were calculated on the basis of the training set part of a corpus of WSJ articles, 1987–1989, consisting of almost 50 million words, which will be described in more detail subsequently.

4.1 Clustering of HC Distributions

In the clustering, n-grams are represented as HC distributions. These distributions are estimated using maximum likelihood as follows:

$$P^r_{w_1w_2}(w_3) = \frac{C(w_1w_2w_3)}{\sum_w C(w_1w_2w)}$$

$$P^r_{w_2}(w_3) = \frac{C(w_2w_3)}{\sum_w C(w_2w)}$$

$$P^r_{\text{UNK}}(w_3) = \frac{C(w_3)}{\sum_w C(w)}$$

$$P^l_{w_3}(w_2) = \frac{C(w_2w_3)}{\sum_w C(ww_3)}$$

$$P^l_{\text{UNK}}(w_2) = \frac{C(w_2)}{\sum_w C(w)}$$

Only a subset of items is clustered. When clustering unigrams we include all 54,243 unigrams that occur more than 10 times in the corpus as well as the unknown word UNK. For mixed clusterings of unigrams and bigrams we include all 378,109 unigrams and bigrams that occur more than 10 times and the unknown word UNK (thus, the unigram set is a subset of the mixed set). We call these sets S_{uni} and S_{mixed} and they are used for all HC and WC models herein, including the Brown model.

We employ bisecting k-means (Steinbach, Karypis, and Kumar 2000) to cluster HC distributions. The distance measure employed is Euclidean distance because the formal properties of k-means, including convergence, only apply to Euclidean spaces. Bisecting k-means is applied to a small random sample of the set of items: k-means first splits this random sample in two, then the largest existing cluster is split and so on until $k = 512$ (or $k = 1{,}024$, depending on the experiment) clusters have been found. The size of the random sample is then doubled, items in the enlarged sample are assigned to cluster

centroids, and centroids are recomputed. The size of the sample is doubled again and so on until all items have been assigned.

Incremental doubling of the sample has the advantage that several iterations of reassignment and recomputation of centroids are performed (thus producing centroids that are good representatives of the overall distribution of items); and that at the same time the total number of assignments that needs to be computed is bounded by $2M$ where M is the number of items. Computing the assignments is responsible for almost all the computation time of k-means and more than 90% of the time needed to estimate the parameters of the exemplar-theoretic model.

We do not investigate the effect of the number of clusters k on the performance of class models in this article. As the default we chose $k = 1{,}024$, similar to Brown et al.'s experiments. Note that we have 512 left HC clusters and 512 right HC clusters, a total of 1,024 in the experiments with $k = 512$. We also experiment with $2 \times 1{,}024$ clusters because one could also argue that this is the setting that is most comparable to Brown et al. We choose the powers of 2, $k = 512$ and $k = 1{,}024$ (instead of 500 and 1,000), for optimal compression and compact storage.

Two examples of half-context clusters (one left HC cluster and one right HC cluster) and their sizes are given in Table 1. The three most frequent words in the left HC cluster have similar left contexts (dominated by forms of *to be*) and different right contexts (large variety of part of speech forms). The three most frequent bigrams in the right HC cluster have similar right contexts (dominated by gerunds) and dissimilar left contexts (again a large variety of possibilities). In traditional whole-context clusters one would need to split the two clusters at least in two. For example, the right HC cluster in Table 1 would have to be split into one subcluster containing *without-first/of-improperly* and one subcluster containing *pain-and*. However, this presents two distinct disadvantages: (1) The extra clusters would require the estimation of more parameters, each based on fewer data points and hence less reliable, and (2) The left-context generalization, whereby *of-improperly* and *pain-and* have similar right contexts, would be lost.

Once clusters and cluster memberships have been computed, we need to determine the relevant right HC cluster c^r_{12} and left HC cluster c^l_3 when computing the probability $P(w_3|w_1w_2)$ according to the model in Figure 1. We do this as follows:

1. If $w_1w_2 \in S_{mixed}$, we use the right HC cluster that $P^r_{w_1w_2}$ was assigned to.
2. Otherwise, if $w_2 \in S_{mixed}$, we use the right HC cluster that $P^r_{w_2}$ was assigned to.
3. Otherwise, we use the right HC cluster that P^r_{UNK} was assigned to.
4. If $w_3 \in S_{uni}$, we use the left HC cluster that $P^l_{w_3}$ was assigned to.
5. Otherwise, we use the left HC cluster that P^l_{UNK} was assigned to.

Table 1
Examples of half-context clusters.

	most frequent n-grams in cluster	size
left HC cluster	unlikely, unclear, happening	753
right HC cluster	pain-and, without-first, of-improperly	248

4.2 Emission and Sequence Probabilities

Emission probabilities need only be estimated for left HC clusters in the exemplar-theoretic model. They are estimated by maximum likelihood:

$$P_e(w|c) = \frac{C(w)}{\sum_{w' \in c} C(w')}$$

Cluster **sequence probabilities** are additively smoothed:

$$P_{\text{seq}}(c^l|c^r) = \frac{C(c^r c^l) + \lambda}{C(c^r) + B\lambda}$$

where $\lambda = 0.1$, $B \in \{512, 1{,}024\}$ is the number of HC clusters, $C(c^r c^l)$ is the number of trigrams $w_1 w_2 w_3$ occurring in the training set, where $w_1 w_2$ was assigned to c^r and w_3 to c^l, and $C(c^r)$ is the number of bigrams $w_1' w_2'$ occurring in the training set, where $w_1' w_2'$ was assigned to c^r.

WC clusters are generated by representing an n-gram as the concatenation of two HC distributions, its left HC distribution and its right HC distribution. Clustering, membership assignment, and probability estimation are the same in all other respects.

5. Experiments and Analysis

A corpus of WSJ articles, 1987–1989, consisting of almost 50 million words, was randomly split into training set (80%), validation set (10%), and test set (10%).

Unigrams, bigrams, and trigrams and their counts were extracted from training, validation, and test sets. A modified KN model (Chen and Goodman 1998), termed $P_{(\text{KN})}$, was estimated on the training set count files and applied to the test set using srilm, the SRI language modeling toolkit (Stolcke 2002). The same count files were the input to the HC and exemplar-theoretic model estimation and application procedure. Vocabulary size was the same for both KN and exemplar-theoretic models: 256,874 (the 256,873 words occurring in the training set and the unknown word). A total of 70.8% of tokens w_3 in the test set occur in a context $w_1 w_2 w_3$ occurring in the training set; for 22.2% of tokens only $w_2 w_3$ occurs in the training set; and for 6.7% only w_3 occurs in the training set. The out-of-vocabulary rate is 0.27%. All validation and test set words that do not occur in the training set are mapped to the special unknown token UNK. In all interpolation experiments, the weight of the $P_{(\text{KN})}$ model is $1 - \alpha$ and the weight of the model with which $P_{(\text{KN})}$ is interpolated is α. The validation set was employed to determine the optimum interpolation weight α and discount D for each case.

Total processing time for estimating the HC clusters for S_{uni} and S_{mixed} (lines 13 and 15 in Table 4, subsequently) was less than 3.5 hours on an Opteron 8214 processor.

In evaluating our model it seems appropriate to compare its performance against other class-based models. Consequently, the SRI toolkit was also used to construct a class bigram language model, following the incremental version of the algorithm proposed by Brown et al. (1992), which we simply term the $P_{(\text{Brown})}$ model. A total of

Table 2
Perplexity results for interpolation of $P_{(\text{Brown})}$ with a bigram model $P_{(\text{KN})}$. $\alpha = 0$ corresponds to $P_{(\text{KN})}$ alone, $\alpha = 1$ corresponds to $P_{(\text{Brown})}$ alone.

	perplexity	
α	validation	test
0.000	164.52	164.80
0.025	164.08	
0.050	164.03	164.33
0.075	164.15	
0.100	164.40	
0.200	166.25	
1.000	245.14	245.45

Table 3
Models used in our experiments.

$P_{(\text{KN})}$	modified Kneser-Ney model
$P_{(\text{ET-Brown})}$	exemplar-theoretic Brown model
$P_{(\text{Half})}$	exemplar-theoretic Half-Context model (Equation 1)
$P_{(\text{Whole})}$	whole-context analogue of Equation 1
$P_{(\text{KN-Brown})}$	interpolation of $P_{(\text{KN})}$ with $P_{(\text{ET-Brown})}$
$P_{(\text{KN-Half})}$	interpolation of $P_{(\text{KN})}$ with $P_{(\text{Half})}$
$P_{(\text{KN-Whole})}$	interpolation of $P_{(\text{KN})}$ with $P_{(\text{Whole})}$

1,024 classes (the same number of classes as the combined left and right context clusters in the 2×512 HC model) were derived from the training data.[1]

Table 2 presents results for the interpolation of $P_{(\text{Brown})}$ with a bigram model $P_{(\text{KN})}$ when applied to the validation set over a number of interpolation weights, followed by results from the test data using the optimum weight for the $P_{(\text{Brown})}$ model ($\alpha = 0.05$) found during the validation phase.

It is clear from Table 2 that although interpolating a traditional class-based model with a KN bigram model does offer some benefit, this benefit is slight (perplexity = 164.80 for $P_{(\text{KN})}$ alone, versus 164.33 using the optimum interpolation weight on the test set). It is also clear that the traditional class-based model operating by itself ($\alpha = 1.0$, perplexity = 245.45) performs poorly relative to $P_{(\text{KN})}$.

Of course the SRI class-based model employs whole-context classes, not half-context distributions which consider behavior to the left and right separately.

The following models, detailed in Table 3, were used in our experiments: modified Kneser-Ney ($P_{(\text{KN})}$); exemplar-theoretic half-context ($P_{(\text{Half})}$); exemplar-theoretic whole-context ($P_{(\text{Whole})}$); exemplar-theoretic Brown ($P_{(\text{ET-Brown})}$); and $P_{(\text{KN-Half})}$, $P_{(\text{KN-Whole})}$, and $P_{(\text{KN-Brown})}$, the interpolations of Kneser-Ney with exemplar-theoretic half-context, whole-context, and Brown, respectively.[2] Perplexity results, for each of these models, from the validation and test sets, are presented in Table 4. Order-2 in Table 4 implies

1 Here we approximate Brown et al. (1992) who used 1,000 classes. As our classes are also being used to investigate language model compression in other work, we prefer to use powers of 2.

2 Test set perplexities for Kneser-Ney in Table 2 (164.80) and Table 4 (165.13, line 1) differ slightly because of different handling of beginning and end of sentence symbols in the two experiments.

Table 4
Perplexity results for Kneser-Ney, exemplar-theoretic Brown, exemplar-theoretic half-context, exemplar-theoretic whole-context, and interpolations.

			perplexity			
	D	α	validation	test	order/model	number of classes
1			164.83	165.13	2 $P_{(\text{KN})}$	
2	.8	1.0	191.71	192.24	2 $P_{(\text{ET-Brown})}$	512
3	1.0	1.0	171.17	171.58	2 $P_{(\text{Half})}$	512
4	1.0	1.0	170.81	171.21	2 $P_{(\text{Whole})}$	512
5	.8	1.0	171.16	171.57	2 $P_{(\text{Half})}$	1,024
6	.8	1.0	170.75	171.19	2 $P_{(\text{Whole})}$	1,024
7	.4	.2	163.97	164.28	2 $P_{(\text{KN-Brown})}$	512
8	.9	.5	161.51	161.82	2 $P_{(\text{KN-Half})}$	512
9	.7	.4	161.83	162.13	2 $P_{(\text{KN-Whole})}$	512
10	.6	.5	161.37	161.67	2 $P_{(\text{KN-Half})}$	1,024
11	.6	.5	161.53	161.83	2 $P_{(\text{KN-Whole})}$	1,024
12			94.67	94.94	3 $P_{(\text{KN})}$	
13	.8	1.0	105.31	105.65	3 $P_{(\text{Half})}$	512
14	.8	1.0	107.99	108.33	3 $P_{(\text{Whole})}$	512
15	.5	.4	88.91	89.15	3 $P_{(\text{KN-Half})}$	512
16	.5	.4	89.39	89.63	3 $P_{(\text{KN-Whole})}$	512

that only unigrams are clustered in the exemplar-theoretic models and the Kneser-Ney model is a bigram model. As for order-3, this implies that both unigrams and bigrams are clustered together in the exemplar-theoretic models and that the Kneser-Ney model is a trigram model.

For lines 7–11 and 15–16, the parameters α and D that were optimal on the validation set are given. For lines 2–6 and 13–14, the optimal value of D on the validation set for $\alpha = 1$ (that is, 0 weight for the $P_{(\text{KN})}$ model) was chosen.

The α parameter on lines 8–11 and 15–16 indicates that half- and whole-context models are as valuable, or nearly so, as the KN models: The interpolation weight of half/whole-context models is either 0.4 or 0.5. In contrast, the Brown class model (line 7) receives a lower weight of 0.2, indicating that it is less valuable in the interpolation with KN.

The discount parameter D determines the influence of class-based generalization in the overall model. Again, the Brown model receives the smallest weight (line 7). For both D and α, the lowest half/whole-context model values are those for the KN order-3 interpolations on lines 15–16: $D = .5, \alpha = .4$ (value of α tied with KN order-2 interpolation on line 9). This may be a reflection of the fact that class-based generalization is contributing more to better performance in order-2 models because order-2 models have a much lower baseline performance.

For order-2 the differences between HC and WC models are small (lines 3 vs. 4, 5 vs. 6, 8 vs. 9, 10 vs. 11). For order-3, exemplar-theoretic half-context is clearly better than exemplar-theoretic whole-context (lines 13 vs. 14), although that difference is reduced in the two interpolated models $P_{(\text{KN-Half})}$ and $P_{(\text{KN-Whole})}$ (lines 15 vs. 16). On this evidence it would appear that the combination of left and right context information into a single context distribution (i.e., a whole-context approach) is redundant, if not harmful. This

is evidence for the half-context hypothesis put forward at the beginning of the article: Outward distributions, present in WC representations but absent in HC representations, do not seem to be helpful in class-based generalization, and are perhaps even harmful in order-3.

One reason why HC models perform better for order-3 than WC models could be that unigrams and bigrams are clustered together for the order-3 models. Although it makes sense to treat, say, the right contexts of *from Mark* and *Martin* as similar, the distributional patterns of the two *n*-grams are very different if the left context is also taken into account, which is the case for WC models.

We could attempt to extend the exchange algorithm that has most often been used for class-based language modeling to half-context clustering. This is beyond the scope of this article, however. Instead, we compare the order-2 WC experiments directly with the Brown classes. We do this to make sure that our good results for HC models are not due to the fact that we use a weak WC baseline. As we will argue now, our WC baseline is at least as good or even better than Brown clustering.

There are two set-ups that can be argued to be directly comparable to the Brown experiments reported here: either 512 left HC classes and 512 right HC classes (lines 2–4, 7–9, and 13–16); or 1,024 left HC classes and 1,024 right HC classes (lines 5–6 and 10–11). In the Brown experiments (lines 2 and 7), Equation (1) is used in the same way as in the HC/WC experiments except that class membership is based on the classes induced by srilm (corresponding to the experiments in Table 2). The comparisons on lines 2 vs. 4 and 6 and 7 vs. 9 and 11 clearly show that the quality of bisecting *k*-means whole-context clustering is comparable to, if not better than, Brown-type whole-context clustering. That is, keeping the representation constant in both cases (i.e., whole-context) enables us to see the algorithmic benefits of bisecting *k*-means as it appears to offer more useful clusters than those produced by the exchange algorithm.

Finally, although the exemplar-theoretic models are clearly outperformed by the $P_{(\mathrm{KN})}$ model (lines 1 vs. 3–6, 12 vs. 13–14), it is important to note that the combination of the $P_{(\mathrm{KN})}$ model and the exemplar-theoretic models outperforms the stand-alone $P_{(\mathrm{KN})}$ model (lines 1 vs. 8–11, 12 vs. 15 and 16). This is strong evidence that a combined class-based and history-length-interpolated model is superior to history-length interpolation by itself.

5.1 Establishing Significance

Although the perplexity results documented here provide tangible support in favor of the half-context hypothesis, it would nevertheless be desirable to establish if the perplexity scores are indicative of improvements that are statistically significant. To this end, the following significance test was performed. The test set has a length of 2,800,613 words. These 2,800,613 positions are divided into 47 bins, corresponding to the part-of-speech of the word at that position that is most frequent in the training set.[3] This was based on a tagging of the training set with TreeTagger (Schmid 1994). One additional bin

3 We initially performed this test by assigning word types randomly to bins. We found that this "random" version of the test was unrealistically sensitive (all differences were highly significant) because differences in perplexity were highly correlated across bins. For example, if a model has a beneficial effect on names only and names are randomly distributed across bins, then perplexity will be better for every bin, which we would then interpret as significance. To reduce correlation across bins, we then defined bins on the basis of part of speech. As a result all names (or rather words whose dominant part of speech is a proper noun) will be in one bin that is not correlated with all other bins.

contains all positions with a number of rare tags (e.g., FW, 'foreign word') and unknown words. Two models are then compared by computing perplexity separately for each bin, counting the number of bins where the first model performs better than the second, and testing the significance of this count using the exact binomial test. This significance test is not very sensitive in some cases because the positive effect of class generalization can be concentrated on a few parts of speech. To the extent that half-context and whole-context classes approximate part-of-speech information, this makes it more difficult to show significance because a number of bins may not be affected by the model. However, as we will see subsequently the test is sufficiently sensitive for the key results of the article.

A number of noteworthy points can be made on the basis of the results of these significance tests. As all models of order-3 significantly (and unsurprisingly) outperform those of order-2, both order types are considered separately in the discussion that follows. All significant results are with respect to p = 0.05.

With regard to the uninterpolated order-2 models (Table 4 models 1–6), the experiments indicate no significant improvement between models, with the exception that all models ($P_{(\text{KN})}$, $P_{(\text{Half})}$, and $P_{(\text{Whole})}$ 512 classes, and $P_{(\text{Half})}$ and $P_{(\text{Whole})}$ 1,024 classes) are significantly better than $P_{(\text{ET-Brown})}$. Although such a result sheds no light on the veracity of the half-context hypothesis, it nevertheless demonstrates that our exemplar-theoretic models are competitive at order-2 and are superior to $P_{(\text{ET-Brown})}$. Concerning the interpolated order-2 models (Table 4 models 7–11), a similar story presents itself, that is, there is no significant difference between the interpolated models with the exception that all interpolated models ($P_{(\text{KN-Half})}$ and $P_{(\text{KN-Whole})}$ 512 classes, and $P_{(\text{KN-Half})}$ and $P_{(\text{KN-Whole})}$ 1,024 classes) improve significantly on $P_{(\text{KN-Brown})}$. It should be noted, however, that the better performance of the five order-2 class-based models (lines 7–11), including $P_{(\text{KN-Brown})}$, compared to $P_{(\text{KN})}$, is statistically significant; this is in keeping with previous findings in the literature and demonstrates that class-based generalization can complement history-length modeling.

As for the order-3 models (Table 4 models 12–16), here the significance results demonstrate that half-contextualization yields statistically significant improvements over whole-context models. Specifically, $P_{(\text{Half})}$ significantly outperforms $P_{(\text{Whole})}$, and $P_{(\text{KN-Half})}$ significantly outperforms $P_{(\text{KN-Whole})}$. In addition, although $P_{(\text{KN})}$ demonstrates superior performance to $P_{(\text{Whole})}$ and $P_{(\text{Half})}$, interpolation with either of our exemplar-theoretic models yields significantly better performance over $P_{(\text{KN})}$ alone. That is, both $P_{(\text{KN-Whole})}$ and $P_{(\text{KN-Half})}$ significantly improve on $P_{(\text{KN})}$.

Overall, these significance results indicate the potential merits of our models. All four exemplar-theoretic models outperform the Brown varieties and the models offer significant improvements versus $P_{(\text{KN})}$ when interpolated at orders 2 and 3. Indeed, $P_{(\text{KN-Half})}$ significantly beats every other model. Crucially, however, in our view, are the results of order-3 which demonstrate the significant benefits of half-contextualization as these lend considerable corroborative weight to our half-context hypothesis.

5.2 Context-Specific Analysis

In order to better understand the relative strengths and weaknesses of the $P_{(\text{KN-Half})}$, $P_{(\text{KN-Whole})}$, and $P_{(\text{KN})}$ models, Table 5 illustrates their performance in fine-grained context-specific detail. $P_{(\text{KN-Half})}$, $P_{(\text{KN-Whole})}$, and $P_{(\text{KN})}$ in Table 5 correspond to lines 15 (order-3 $P_{(\text{KN-Half})}$), 16 (order-3 $P_{(\text{KN-Whole})}$), and 12 (order-3 $P_{(\text{KN})}$) in Table 4, respectively.

The table is a stratification into 17 strata of the positions w_3, occurring in context $w_1w_2w_3$, in the validation set according to length $|h|$ of history used by the half-context

Table 5
Context-specific analysis of model performance. $|h|$ is the length of the history used by $P_{\text{(KN-Half)}}$ and $P_{\text{(KN-Whole)}}$ for prediction. f_3 is the training set frequency of w_3. $f_{1,3}$ is the training set frequency of $w_1w_2w_3$. The number of tokens and types of $w_1w_2w_3$ for the validation set are also provided. KN-Half corresponds to line 15 in Table 4 (the interpolation of Kneser-Ney and exemplar-theoretic half-context); KN-Whole to line 16 (the interpolation of Kneser-Ney and exemplar-theoretic whole-context); and KN corresponds to line 12 (the order-3 $P_{\text{(KN)}}$ model); . $L_{\text{(KN-Half)}}$ is the log likelihood of $P_{\text{(KN-Half)}}$ on the validation set. $\Delta_{\text{(KN-Whole)}}$ and $\Delta_{\text{(KN)}}$ are the differences in log likelihood of these models from $L_{\text{(KN-Half)}}$ on the validation set. The l value gives average per-position log likelihood on the validation set and δ values per position differences. The three largest absolute differences in each delta column are in **bold**.

	$\lvert h\rvert$	f_3	$f_{1,3}$	tokens	types	$L_{\text{(KN-Half)}}$	$\Delta_{\text{(KN-Whole)}}$	$\Delta_{\text{(KN)}}$	$l_{\text{(KN-Half)}}$	$\delta_{\text{(KN-Whole)}}$	$\delta_{\text{(KN)}}$
1	0	≥0	≥0	62,810	61,453	−151,604	−4	1,545	−2.41	−0.00	0.02
2	1	0–9	0	10,529	10,482	−54,333	−17	−1,770	−5.16	−0.00	−0.17
3	1	1–9	≥1	1,249	1,132	−2,837	0	−114	−2.27	0.00	−0.09
4	1	≥10	0	673,816	666,307	−1,626,791	−114	**24,439**	−2.41	−0.00	0.04
5	1	≥10	1	115,562	109,492	−170,263	139	10,378	−1.47	0.00	0.09
6	1	≥10	2	48,165	43,296	−34,540	31	−2,385	−0.72	0.00	−0.05
7	1	≥10	3–4	42,903	35,839	−21,703	18	−3,474	−0.51	0.00	−0.08
8	1	≥10	5–9	33,199	23,429	−9,045	8	−3,123	−0.27	0.00	−0.09
9	1	≥10	≥10	1,967	1,153	−282	0	−158	−0.14	0.00	−0.08
10	2	0–9	0	33,410	33,076	−191,114	816	−5,866	−5.72	**0.02**	**−0.18**
11	2	1–9	≥1	9,278	8,310	−42,757	1	4,479	−4.61	0.00	**0.48**
12	2	≥10	0	718,269	702,377	−2,891,127	**5266**	**−72,624**	−4.03	**0.01**	−0.10
13	2	≥10	1	259,856	241,428	−724,464	**1324**	**109,446**	−2.79	**0.01**	**0.42**
14	2	≥10	2	161,628	141,797	−383,855	717	22,986	−2.37	0.00	0.14
15	2	≥10	3–4	214,073	173,784	−447,205	693	19,603	−2.09	0.00	0.09
16	2	≥10	5–9	308,716	211,119	−553,419	745	13,821	−1.79	0.00	0.04
17	2	≥10	≥10	2,407,403	337,259	−2,639,923	**2179**	21,844	−1.10	0.00	0.01

and whole-context models (0, 1, or 2), frequency f_3 of w_3 in the training set, and frequency $f_{1,3}$ of $w_1w_2w_3$ in the training set. Each line gives statistics for one stratum. We explain the statistics for the example of stratum 13. This stratum contains all positions w_3 in the validation set that satisfy the following three conditions: w_1w_2 is a bigram that half- and whole-context models use for class-based prediction ($|h| = 2$); w_3's frequency in the training set is at least 10; and the trigram $w_1w_2w_3$ occurred exactly once in the training set. There are 259,856 validation set positions in this stratum, corresponding to 241,428 different trigram types $w_1w_2w_3$. The log likelihood of this subset of the validation set for $P_{\text{(KN-Half)}}$ is −724,464. This log likelihood of −724,464 is better by 1,324 than that of $P_{\text{(KN-Whole)}}$ and better by 109,446 that that of $P_{\text{(KN)}}$. The per-position (average) log likelihood of $P_{\text{(KN-Half)}}$ is −2.79. This per-position log likelihood is better by 0.01 than that of $P_{\text{(KN-Whole)}}$ and better by 0.42 that that of $P_{\text{(KN)}}$.

The 17 strata were chosen so as to get good resolution on the contexts that distinguish the models. These are the contexts that contain a history that is used by the class models (lines 4–9 and 12–17). The other five strata (1, 2, 3, 10, 11) are comparatively small and have a small impact on overall difference in log likelihood. The KN model interpolates predictions for histories of different lengths. In general, this will include the

history that the class-based models use, but also include other lengths. For example, in cases where the class-based model is using a length-2 history, the KN model interpolates length-2, length-1, and length-0 histories.

We first compare $P_{\text{(KN-Half)}}$ and $P_{\text{(KN)}}$. In general, $P_{\text{(KN-Half)}}$ performs better than $P_{\text{(KN)}}$ if a history of length 2 is available for prediction and the predictee w_3 occurred at least once in the identical context in the training set (lines 11, 13–17, $\Delta_{\text{(KN)}}$ is positive for these strata). Stratum 11 contains a small number of positions and consequently contributes little to $\Delta_{\text{(KN)}}$, but the per position difference is the largest of any stratum (0.48). In contrast, the per-position difference for stratum 17 is small, but the overall contribution to $\Delta_{\text{(KN)}}$ is still noticeable since this stratum is the largest. The overall contribution of all length-2 history strata to $\Delta_{\text{(KN)}}$ is 113,689 (the sum of rows 10–17 of the corresponding column) and is thus responsible for the majority of the perplexity improvement due to half-context modeling.

The two strata 10 and 12 are exceptions. Per-position decrease in performance is large for half-context modeling and the overall impact is also large for stratum 12. In these two cases, $P_{\text{(KN)}}$ backs off to context length 1 for prediction. Recall that, in contrast, $P_{\text{(KN-Half)}}$ uses only one context length for prediction, the longest that is available. So on line 12, $P_{\text{(KN-Half)}}$ underpredicts the next word w_3 using a bigram w_1w_2 because w_3 did not occur in that position in the training set ($f_{1,3} = 0$). $P_{\text{(KN)}}$ can also use the unigram w_2 for prediction and computes better estimates in cases where w_2w_3 occurred in the training set. We are planning to address this problem in future work on the half-context model.

A similar problem for $P_{\text{(KN-Half)}}$ can be observed on lines 6–9. In these cases, $P_{\text{(KN)}}$ can use both the unigram w_2 and the bigram w_1w_2 for prediction. Note that the values for $f_{1,3}$ indicate that the trigram $w_1w_2w_3$ occurred at least twice in the training set. $P_{\text{(KN-Half)}}$ only uses the unigram w_2 because the bigram w_1w_2 was not frequent enough to be included in the model as a bigram. The results on lines 6–9 suggest that further improvements of $P_{\text{(KN-Half)}}$ are possible by interpolating predictions of histories of different lengths.

Large improvements are realized by $P_{\text{(KN-Half)}}$ in strata 4 and 5. For these two strata, the trigram $w_1w_2w_3$ has frequency 0 or 1 and therefore the length-2 component of $P_{\text{(KN)}}$ does not predict w_3 well. $P_{\text{(KN-Half)}}$ achieves good predictions because the single word w_2 used for prediction occurred frequently ($f_3 \geq 10$) and its class therefore is likely to reflect the distributional properties of w_2 well.

In summary, $P_{\text{(KN-Half)}}$ is the overall superior model because it successfully employs class-based generalization for rare events. However, for a number of strata (6–9, 10, 12) $P_{\text{(KN-Half)}}$ only uses one context for prediction, which in many cases is an inferior choice compared to the predicting contexts used by $P_{\text{(KN)}}$. As a result, the averages in these strata are negative. We plan to address this problem in future work (see Section 6).

Turning to the differences between $P_{\text{(KN-Half)}}$ and $P_{\text{(KN-Whole)}}$, we see that these differences are quite small—some positive, some negative—for length-1 histories (strata 2–9). Only for length-2 histories do we find larger differences: strata 12, 13, and 17 for total differences on the validation set and strata 10, 12, and 13 for average differences. Length-2 differences are consistently positive for $P_{\text{(KN-Half)}}$ although some of the differences are small.

The better relative performance of $P_{\text{(KN-Half)}}$ for length-2 histories can be explained by the fact that the difference between half-context and whole-context models increases as the size of n-grams being clustered grows. For unigrams, there is significant correlation between left and right half-contexts: If two words have the same type of right context, then they often also have the same type of left context. For bigrams, this correlation is smaller. As an illustration consider the bigram *underwriter was*.

It occurs four times in the validation set, followed by the words *Dillon*, *Hambrecht*, *Merrill*, and *Nesbitt*, none of which occur in this context in the training set. For all four words $\log[P_{\text{(KN-Half)}}(w_3|\text{underwriter was})/P_{\text{(KN-Whole)}}(w_3|\text{underwriter was})] \approx 2$, that is, $P_{\text{(KN-Half)}}$ has a large advantage compared to $P_{\text{(KN-Whole)}}$. The reason is that *underwriter was* is in the same right half-context bigram class as many bigrams that are followed by *Dillon*, *Hambrecht*, *Merrill*, or *Nesbitt* in the training set. Examples of such bigrams include *banking at*, *bonds via*, *firm of*, *sold through*, and *strategist at*. These bigrams have similar right contexts, but dissimilar left contexts. As a result, the whole-context model groups *underwriter was* with other bigrams that do not support good generalization. The pattern of consistent improvements of $P_{\text{(KN-Half)}}$ compared to $P_{\text{(KN-Whole)}}$ for bigrams (lines 11–17) indicates that half-context clustering is able to capture useful generalization for language modeling that whole-context clustering cannot capture.

For unigrams, there are many cases that show the same effect. For example, the unigram *persuades* is followed by *them* in the validation set, again a context unseen in the training set. The half-context model groups *persuades* with *n*-grams like *They told*, *and prevented*, and *both of*—dissimilar on the left, but similar on the right—that support a high estimate for $P_{\text{(KN-Half)}}(\text{them}|\text{persuade})$. The class of *persuade* of the whole-context model is more diffuse, generally containing *n*-grams that are followed by a noun phrase, but in contexts like *act until*, *admit that*, *clearance by* that make a following *them* less likely.

However, there are also unigrams where class membership in the whole-context model leads to better generalization than class memberships in the half-context model because the whole-context model can exploit the correlation of left and right contexts. For example, the whole-context model assigns the unigram *367,000* to a class that consists almost exclusively of numbers whereas the half-context model assigns it to a class that is more mixed. Because *367,000* occurs in only 11 distinct contexts in the training set, its syntactic behavior can be better characterized if both left and right contexts are exploited.

In summary, half-context and whole-context models perform similarly on average for length-1 histories although there are large differences between the two models for individual 1-word histories. For length-2 histories, the half-context model is superior due to its ability to group histories according to the relevant half-context only—the right half-context—in accordance with the half-context hypothesis.

5.3 Objective Function of *n*-gram Clustering

As we argued in Section 3.3 when introducing the exemplar-theoretic model, class-based generalization is most useful for unseen and for infrequent events. This basic insight motivates two differences between our class-based model and previous work.

First, the discounting mechanism defined in Equation (1) varies the weight that class-based generalization is given: Weights for unseen and infrequent events are higher than weights for frequent events. As a result, the model's estimates are close to maximum likelihood estimates for frequent events because the maximum likelihood estimator is appropriate in these cases. In contrast, the model's estimate of the probability of a word occurring in an unattested context is closer to the estimate of the class-based model.

Using class-based generalization only for rare events also has implications for the objective function of clustering. Most previous work has employed objective functions that optimize a quantity on the *entire* training set. For example, Brown et al. (1992) maximize mutual information and Gao et al. (2002) minimize perplexity on the entire

training set. In contrast, the objective function of our clustering is similarity of half-contexts to cluster centroids or, more precisely, minimizing the residual sum of squares of differences between half-context vectors and cluster centroids. This criterion is much less sensitive to frequency than previously used criteria. In the extreme case, it may be optimal on the global criteria to put a very frequent idiosyncratic word in its own class. This is so because even slightly better improvements of the class model for a frequent word will affect many positions in the training set and have a large cumulative effect.

Our objective function is not influenced by frequency because vectors are normalized. In principle, the gain from finding an appropriate cluster for a rare word is as large as the gain from finding an appropriate cluster for a frequent word. In particular, frequent idiosyncratic words have no advantage compared to infrequent idiosyncratic words and very frequent idiosyncratic words are less likely to be assigned to singleton clusters or small clusters dominated by them. This clustering set-up may not be optimal for achieving good results with a cluster-only language model, that is, a model that does not contain a "lexical" component similar to the maximum likelihood estimates in our exemplar-theoretic model. But if we acknowledge that class-based generalization is not useful or is even harmful for frequent events, then this should not be our goal.

In summary, we attribute part of the success of our half-context models to the fact that both the design of the discounting mechanism and the *k*-means objective function target a different subspace of the space of all events: those that are unseen or infrequent.

5.4 Efficient Clustering

The focus of this article is the comparison of HC and WC classes and our investigations into the context-specific characteristics of history-length interpolation and class-based generalization. However, we also want to point out that the clustering algorithm we are using is very efficient, thus removing a potential obstacle to the widespread use of class-based language models. In total, the clustering algorithm requires fewer than two assignments per item on average (see Section 4.1). A single assignment requires computing the distance between an HC distribution and each of k centroids. The time necessary for computing one distance is a function of the number of nonzero entries in the distribution.[4] The total number of nonzero entries for any given bigram is the number of distinct trigrams in which it occurs. Thus, the total number of operations for performing all assignments necessary for the clustering of the bigrams is less than b times the number of distinct trigrams in the corpus where b is a small constant. This number scales linearly with the number of distinct trigrams, which in turn scales sub-linearly with the length of the training corpus. Thus, although the estimation procedure is expensive compared to standard trigram models like KN, it has desirable properties compared to other clustering algorithms, in particular the exchange algorithm. Even though there exist fast implementations for the exchange algorithm (Martin, Liermann, and Ney 1998; Uszkoreit and Brants 2008), it has worse than linear complexity.

6. Conclusions and Future Work

In this article we introduced a new representational formalism for language modeling known as *half-contextualization*. Half-contextualization employs only *inward* contextual

4 This only holds for the dot product, not for the Euclidean distance, but the latter can be computed from the former as $\sum(x_i - y_i)^2 = \sum x_i^2 + \sum y_i^2 - 2\sum x_i y_i$ if sums of squares are precomputed and cached.

information in estimation and prediction—where we defined the inward distributions as the conditioning context's right-context distribution and the predicted word's left-context distribution. Our hypothesis was that only inward context is helpful for accurate prediction.

The experimental results and statistical analyses herein indicate that this hypothesis is correct and that the use of outward directed information is not only redundant but also, in the case of order-3, damaging. We believe this is a particularly noteworthy discovery as it is essentially tantamount to requiring only half of the available distributional information in order to achieve an equivalent, and often better, result. Furthermore, from the outset we argued that the lack of adoption of class- and similarity-based approaches was, in part, because the granularity of contexts best suited for generalization and history-length interpolation have yet to be established; the novel context-specific analysis we presented here, which goes beyond traditional perplexity comparisons between models, illustrates the specific context scenarios where half-contextualization is particularly beneficial.

The HC hypothesis is at first counter-intuitive: Standard language models treat words as atomic units that are best characterized by taking into account all information available about them in the training set, including what we call outward context. The model we have proposed uses different parts of the available contextual information for different inference tasks. While it may seem surprising that contextual information can be redundant or harmful for class-based generalization, we have argued that directionally nonrelevant information for a particular inference task can be noisy and misleading.

In addition to half-contextualization, we introduced three other innovations for class-based language models. First, we defined classes as mixed classes of bigrams and unigrams and argued that this flexible granularity gives rise to better classes. Second, we successfully employed a discounting method which focuses the impact of generalization onto rare events while leaving frequent events to better-suited history-length interpolation. This addresses the problem that class-based generalization is often harmful for high frequency events that are best estimated by maximum likelihood on identical contexts. Third, we presented a new clustering algorithm for class-based language models that has linear time complexity and is more efficient than the exchange algorithm.

With regard to the future development of our exemplar-theoretic model, one obvious avenue, given our analyses, is to incorporate the ability to interpolate distributions of different-length conditioning contexts into the model. By incorporating such an interpolation mechanism, we anticipate an amelioration in performance further supporting the use of half-context in language models. However, crucially, the primary endeavor in this article is not simply to promote the merits of half-contextualization, nor to establish how to build a better exemplar-theoretic model, but rather to develop and promote a deeper understanding of the relationship between history-length interpolation, class-based generalization, and context, in order to construct and combine language models, of varying varieties, in a more targeted fashion.

Acknowledgments

The research presented in this article was funded by DFG grant SFB 732. We would like to thank Helmut Schmid and audiences at Google Mountain View research, the Berkeley International Computer Science Institute, the 2010 Google EMEA faculty summit, and the reviewers for their constructive comments.

References

Abdoos, Monireh and Seyed Gholamreza Jalali Naeini. 2008. Word categorization using clustering ensemble. In *Proceedings of the 2008 International Conference on Advanced Computer Theory and Engineering*, pages 662–666, Washington, DC.

Bahrani, Mohammad, Hossein Sameti, Nazila Hafezi, and Saeedeh Momtazi. 2008. A new word clustering method for building n-gram language models in continuous speech recognition systems. In *Proceedings of the 21st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems: New Frontiers in Applied Artificial Intelligence*, IEA/AIE '08, pages 286–293, Berlin.

Bai, Shuanghu, Haizhou Li, Zhiwei Lin, and Baosheng Yuan. 1998. Building class-based language models with contextual statistics. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 173–176, Seattle, WA.

Bassiou, Nikoletta and Constantine Kotropoulos. 2011. Long distance bigram models applied to word clustering. *Pattern Recognition*, 44:145–158.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Brown, Peter F., Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Chen, Stanley F. and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report TR-10-98, Harvard University, Cambridge, MA.

Dagan, Ido, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

Emami, Ahmad and Fred Jelinek. 2005. Random clustering for language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages I:581–584, Philadelphia, PA.

Essen, Ute and Volker Steinbiss. 1992. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages I:161–164, San Francisco, CA.

Gao, Jianfeng, Joshua T. Goodman, Guihong Cao, and Hang Li. 2002. Exploring asymmetric clustering for statistical language modeling. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 183–190, Morristown, NJ.

Hall, Keith and Mark Johnson. 2003. Language modelling using efficient best-first bottom-up parsing. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 507–512, St. Thomas.

Hintzman, Douglas L. 1986. 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, 93:328–338.

Justo, Raquel and M. Inés Torres. 2009. Phrase classes in two-level language models for ASR. *Pattern Analysis & Applications*, 12(4):427–437.

Kneser, Reinhard and Hermann Ney. 1993. Improved clustering techniques for class-based statistical language modelling. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 973–976, Berlin.

Martin, Sven, Jörg Liermann, and Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech Communication*, 24:19–37.

Momtazi, Saeedeh, Sanjeev Khudanpur, and Dietrich Klakow. 2010. A comparative study of word co-occurrence for term clustering in language model-based sentence retrieval. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 325–328, Morristown, NJ.

Momtazi, Saeedeh and Dietrich Klakow. 2009. A word clustering approach for language model-based sentence retrieval in question answering systems. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1911–1914, Hong Kong.

Ney, Hermann, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–28.

Nosofsky, Robert M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper, editors, *Frequency Effects and the Emergence of Lexical Structure*. John Benjamins, Amsterdam, pages 137–157.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision

trees. In *Proceedings of Conference on New Methods in Language Processing*, pages 44–49, Manchester.
Schütze, Hinrich. 1993. Distributed syntactic representations with an application to part-of-speech tagging. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 1504–1509, San Francisco, CA.
Schütze, Hinrich. 1995. Distributional part-of-speech tagging. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 141–148, Belfield.
Schütze, Hinrich and Michael Walsh. 2008. A graph-theoretic model of lexical syntactic acquisition. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–926, Honolulu, HI.
Schwenk, Holger and Philipp Koehn. 2008. Large and diverse language models for statistical machine translation. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 661–666, Hyderabad.
Steinbach, Michael, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. Paper presented at the Workshop on Text Mining, Knowledge Discovery and Data Mining, Boston, MA.
Stolcke, Andreas. 2002. Srilm—an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, CO.
Uszkoreit, Jakob and Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 755–762, Columbus, OH.
Wiegand, Michael and Dietrich Klakow. 2008. Optimizing language models for polarity classification. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 612–616, Berlin.
Zitouni, Imed. 2007. Backoff hierarchical class n-gram language models: Effectiveness to model unseen events in speech recognition. *Journal of Computer Speech and Language*, 21(1):88–104.
Zitouni, Imed and Qiru Zhou. 2007. Linearly interpolated hierarchical n-gram language models for speech recognition engines. In Michael Grimm and Kristian Kroschel, editors, *Robust Speech Recognition and Understanding*. InTech, Vienna, pages 301–318.
Zitouni, Imed and Qiru Zhou. 2008. Hierarchical linear discounting class n-gram language models: A multilevel class hierarchy approach. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4917–4920, Las Vegas, NV.