

# Occurrence Based Statistics in Machine Translation

**Vamshi Ambati**

Language Technologies Institute  
Carnegie Mellon University  
vamshi@cs.cmu.edu

**Alon Lavie**

Language Technologies Institute  
Carnegie Mellon University  
alavie@cs.cmu.edu

## Abstract

As MT approaches demand longer context for better translation quality, the limitations of current language modeling techniques become explicit. The computational inability to model the likelihood of longer ngrams and the likelihood of their usage in probabilistic manner, have prevented us from exploring long ngrams in MT. In this paper, we propose and investigate a new set of features called occurrence based statistics for machine translation that overcome these limitations. Occurrence based statistics are obtained by looking up the existence of sequence of lexical units (ngrams) in huge human created language repositories without actually dealing with their probabilities. We also experiment occurrence based statistics for certain syntactic units like the headword chains (nchains). Our experiments show that they correlate well with translation quality and are useful as discriminative features in classification too.

## 1 Introduction

Present day Statistical Machine Translation systems make extensive use of a Language Model, built over a very large corpus of the language. In particular Language

Modeling is extensively used in the decoding phase of Machine Translation systems for the task of distinguishing between well formed fluent translation hypotheses from the rest. Although language models are effective in handling disfluencies in translation that are caused due to local reordering of lexical items, they fail miserably when working with highly divergent language pairs with global lexical reordering. The main reason being that most LMs used for MT are restricted to trigram language models due to computational feasibility.

Some MT groups are investigating the feasibility of improving the capacity of language models by accommodating longer ngrams (Zhang and Vogel, 2005) (Callison-Burch et al., 2005). Although this is promising, estimating the probabilities with these LMs of long phrases requires vast computational resources. Google, for example, uses an immense distributed computer farm to work with 6-gram LMs. Many researchers will not be able to afford to compute these LMs or at least use them in their systems.

Even with long ngrams statistical LMs it is not completely clear, how we would overcome the problem of low likelihood of long ngrams. For example during the decoding phase a hypothesis that has high probability calculated from individual smaller units of ngrams is favored over other hypotheses, although consisting of long ngrams with low likelihood probability. Studies show that there are a lot of sentences quite below in the nbest lists

formed as intermediate outputs during the decoding, which have better hypotheses than the highly likely chosen one and reranking of the nbest hypotheses is frequently performed to improve translation quality (Shen et al., 2004).

This has motivated us to look at the problem of language modeling as merely a discriminative method that returns a binary score of the validity of a sequence of lexical units. We propose that long sequences of words can be used as discriminative features without directly modeling their probability, by determining whether a given sequence of  $n$  words has ever been seen in large corpora of raw human-authored text. This can be done using IR methods against vast monolingual text collections, and ultimately, it may be computationally feasible to execute such queries over the entire web, or a local snapshot of the web. These features are what we call "occurrence based statistics". Occurrence based statistics are obtained by looking up the existence of sequence of lexical units (ngrams) in huge human created language repositories without actually dealing with their probabilities. We also experiment occurrence based statistics for certain syntactic units like the headword chains (nchains).

The rest of the paper is organized as follows. In section 2 we discuss some related work that motivates and is related to our approach. Section 3 discusses language modeling and issues when applied in Machine Translation. Section 4 discusses the proposed set of features in details. Section 5 discusses the framework for extracting such features on a large scale. Section 6 discusses how syntax features can be introduced in our framework. Section 7 discusses the experiments and results in detail. We conclude by discussions and future work.

## 2 Related Work

To our knowledge an investigation of occurrence based statistics for machine translation has not been pursued so far and there is relatively less prior work. Web-ngrams proposed by Kevin Knight (Koehn and Knight, 2003)

have been used for NP translation. In a different scenario, Dan Gileidia and others explored the discriminative nature of syntactic features like headword chains in the context of MT Evaluation. We propose occurrence based features of ngrams and other syntactic units as a more convenient and effective way for modeling language.

## 3 Language Modeling and Problems for MT

Language modeling aims to capture language as a probabilistic distribution and uses it elsewhere to distinguish between a probable sentence in the language from a very improbable one. Statistical approaches to MT have relied on language models to prune the gigantic search space encountered in the process of decoding by maximizing the probability of the final target translation based on pre-computed probabilistic distributions.

The language models used by most if not all of today's state-of-the-art MT systems are traditional statistical ngram models. These LMs were originally developed within the speech recognition research community. MT researchers later adopted these LMs for their systems, often as is. The CMU-Cambridge-SLM and SRI-LM toolkits, originally developed for speech, are broadly used to construct the LMs used in most current MT systems. While there has been a large volume of language modeling work within the speech community, it has often been observed that sophisticated LMs are unable to outperform simple appropriately smoothed trigram LMs. Also the phrase based machine translation (Koehn et al) have shown that the likelihood of longer ngrams or phrases being used in the translation is quite low and so the computational complexity and space complexity outweighs the improvement in translation quality due to long ngrams. Consequently, trigram LMs are the most commonly used, not only in speech, but also in MT. Although the characteristics of the search space in MT is fundamentally different than in speech, there has not been much prior work in investigating the suitability of using trigram models for

MT. Whereas in speech, competing alternatives are usually homonyms, with little semantic relationship, and the time-linear ordering of the words results in hypotheses that often follow basic syntax, the translation alternatives in MT are often semantically related, and divergent syntax between source and target languages often results in highly mis-ordered ungrammatical and disfluent hypotheses.

#### 4 Occurrence Based Statistics as Features

Traditional statistical language models aim to assign a probability to a given sequence of words that is reflective of its relative likelihood in the language. Reliably estimating such probabilities however is a challenging task. LMs that use longer ngram histories hope to derive better estimates, but face harsh issues of training data sparsity, and require vast computational resources for storing and retrieving the ngram counts and probabilities. The basic MT search scenario, however, does not inherently require exact probabilities of strings. What is needed is a source of information that can reliably identify word sequences that are grammatical in the target language, and that can distinguish such sequences from among the large number of alternative word sequences hypothesized by the MT system that are ungrammatical.

This observation inspired us to look at the problem differently. Sequences of words that have been observed to occur in human-authored texts can be assumed to be grammatical by nature. On the other hand, the grammaticality of word sequences that cannot be found in human-produced texts is unknown. It is possible that they are ungrammatical. Its also possible that a human simply never produced this exact sequence. Poor translations produced by MT systems commonly violate correct word and phrase order. Consequently, many word sequences from such translations are ungrammatical, and will not have occurred in human-authored text. More grammatical translations, however, will contain more grammatical word se-

quences that are therefore more likely to be found in human-authored text. This bias suggests that occurrence statistics for reasonably long word (ngram) sequences can in fact serve as good discriminators between good and bad translations. ngrams that are too long, however, are unlikely to have been observed, and ngrams that are too short may not discriminate effectively. We conjecture that there exists a middle range of ngram orders which can provide us with discriminative information.

In general, ngram level occurrence based statistics can be computed as -

$$OBS(n) = \sum_{k \in S} \frac{f(k)}{|S|}$$

$$S = \{units\ of\ size\ n\}$$

$$f(k) = \begin{cases} 1 & \text{if present in corpus} \\ 0 & \text{otherwise} \end{cases}$$

The metrics that we propose in this paper are occurrence based statistics of sub-sentential and sub-structural elements in a huge monolingual corpus. These metrics are inspired from the roots of MT Evaluation research, where a sentence is judged of high translation quality by the occurrence of fragments in the reference human translation. In fact the state of the art evaluation approaches use intelligent ways to calculate these overlaps by using syntactic measures (Liu and Gildea, 2005). When such reference translations are not available.

#### 5 Infrastructure for Computing OBS Features

The basic phases in computing occurrence statistics are two. First is the construction of large databases of feature statistics that are acquired and calculated over large amounts of grammatical human-authored monolingual data. Second are fast and efficient techniques to perform lookup to obtain the statistics. In this section we discuss in detail each of the phases.

## 5.1 Data Creation

With the advent of various Language initiatives on the Internet, we find a plethora of corpus available for various languages. For some languages, there is more corpus that can be handled by computers at this point in time. The bottleneck seems to be the apparatus and algorithms used by systems. There is about 13TB word tokens on the internet, but the largest possible Language model built and used to describe English has only been 5 TB of data. Even though the data is publicly available, it is still difficult for researchers to make benefit of it due to the lack of computational resources. Only large companies like "Google" have been . With intelligent algorithms for language modeling (SALM toolkit) major strides have been made, but we are still far away from benefitting from the data available to us.

Universal Library Project <sup>1</sup> at Carnegie Mellon University is another rich source of human authored content. The aim of the project is to accumulate in digital format all the human knowledge so far documented in print media. It has a huge collection of literature from various languages. There are currently 200 thousand English books alone and many books belonging to different languages. The text content in it as opposed to the WWW is more grammatical and formal. Such resources can better be exploited to obtain reliable occurrence based statistics.

## 5.2 Lookup Techniques

The lookup function to retrieve the statistics need to be quick and fast.

### 5.2.1 Suffix Array Lookup Technique

The suffix array data structure (Manber and Myers, 1990) was introduced as a space-economical way of creating an index for string searches. The suffix array data structure makes it convenient to compute the frequency and location of any substring or ngram in a large corpus. Abstractly, a suffix array is an alphabetically-sorted list of all suffixes in a corpus, where a suffix is a substring running

from each position in the text to the end. However, rather than actually storing all suffixes, a suffix array can be constructed by creating a list of references to each of the suffixes in a corpus. Figure 1 shows how a suffix array is initialized for a corpus with one sentence. Each index of a word in the corpus has a corresponding place in the suffix array, which is identical in length to the corpus. This is further sorted to form a list of the indices of words in the corpus that corresponds to an alphabetically sorted list of the suffixes.

Suffix arrays have been used as a super-efficient data structure for scaling phrase based SMT systems (Callison-Burch et al., 2005). For our purpose we only use the suffix array as an efficient string retrieval mechanism.

### 5.2.2 Traditional Information Retrieval Technique

Although Suffix Arrays is an effective data structure for storing and handling large corpora, the space complexity might quickly become unmanageable when scaling to the terabyte or petabyte corpora that we might want to use in future. Since better occurrence based statistics can be obtained by looking at large volumes of corpora. The other concern is that it is not completely straightforward to incorporate syntactic information or extend the occurrence based statistics to handle generalizations at various level.

For this we explore the feasibility of traditional Information Retrieval methods that already are famous for scaling to terabytes of data at economical costs of complexity. Also recently some IR groups have been looking at structured query languages for Information Retrieval task. Indri search engine (Metzler et al., 2004) built on top of the Lemur toolkit <sup>2</sup> has been successful in indexing and storing annotated data as well as providing sophisticated retrieval languages that exploit these syntactic annotations. For example Indri search engine facilitates queries like:

---

<sup>1</sup><http://www.ulib.org>

---

<sup>2</sup>[www.lemurproject.org](http://www.lemurproject.org)

*The #any:country ambassador will be visiting #any:country in december*

*This is a #any:ADJ place*

This will enable obtaining better occurrence based statistics on sparse data sets without loss of generality. This could easily be extended to handle any level of generalizations either at syntactic categories like named entities, part of speech etc or semantic categories.

## 6 Syntax based Occurrence Statistics

In addition to the information that can be gathered from surface words, we hypothesize that statistics about deeper linguistics phenomena gathered at the syntactic and semantic level also have discriminative powers. The idea discussed in section 2.1 can be naturally extended to deeper linguistic levels. For example, instead of looking for the occurrence of a long ngram sequence (from a translation candidate) in a human-produced corpus, we might look for the occurrence of a parse-tree fragment. If we have never seen a particular syntactic construct proposed by a translation candidate in a large treebank of well-formed English sentences, there is a good chance that the proposed construction is not grammatical.

The potential benefit of using linguistic features as discriminators is that they are a more restrictive model of grammaticality judgments than word statistics (i.e., in addition to co-occurrence, words must form the right relationship). A major challenge is that this approach faces a more serious data sparsity problem. Another challenge is that most off-the-shelf NLP applications are not intended to process ungrammatical sentences such as those produced by translation systems. With respect to the first challenge, we focus on capturing head-modifier relationships with our features instead of full argument frames. With respect to the second challenge, we perform direct structural comparisons between a translation sentence against

a model of well-formed English sentences instead of using parse tree probabilities as features. Similar to the argument made earlier for matching long ngrams instead of trusting standard language modeling scores, we believe that features that verify the existence of parse fragments will be more indicative of the grammaticality of the translation sentence.

### 6.1 Headword Chains

A headword chain is a sequence of words which corresponds to a path in the dependency tree. In order to extract headword chains we first parsed the sentences using a dependency parser (McDonald et al., 2005) to obtain a dependency tree. From the dependency tree, all possible headword chains of different orders were extracted which is basically all possible paths starting from any node in the tree to the root node. The original distribution of the parser is trained only on a subset of the Penn Tree Bank (Marcus et al., 1994) and so we retrained the parser on the entire Penn Tree Bank corpus.

Optionally, a full constituent parser can also be used to obtain a syntactic tree and then obtain a dependency tree from it. Dependency trees can be obtain from syntactic constituent trees by applying the deterministic headword extraction rules used by the Yamada and Matsumoto (Yamada and Matsumoto, 2003) or other available tools like Penn2Malt <sup>3</sup>

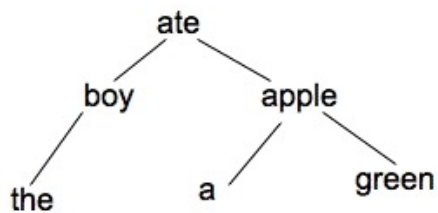


Figure 1: Dependency Tree

The dependency tree is a kind of structure constituted by headwords and the top node is the root headword and every subtree represents the modifier information for its root

<sup>3</sup><http://w3.msi.vxu.se/nivre/research/Penn2Malt.html>

headword. For example, the dependency tree in the syntax tree in Figure 1 corresponds to a sentence "The boy ate an green apple". The ngrams corresponding to order '1' are all the unigrams or the individual words. The 2grams in the sentence are "the boy", "ate an", "green apple", "boy ate" and "an green". Similarly the higher order ngrams can be enumerated. The maximum order of ngrams in this sentence is a 6gram which is the entire sentence.

Similarly for the same example the headword chain information is as follows. To draw correspondence easily we name these as nchains. So the 1chain are all the individual headwords which is all the nodes in the dependency tree. A 2chain is a sequence of words at any two nodes - a parent and a sibling. The 2chains in the above example are "the boy", "boy ate", "a apple", "green apple", "apple ate". Similarly the 3chains are "the boy ate", "a apple ate", "green apple ate". There are no 4chains in this particular example as the depth of the tree is only 3.

## 7 Experiments

### 7.1 Data and Setup

For the data repository we collected the source side of the Europarl corpus and indexed it using the SALM toolkit (Zhang and Vogel, 2006). We only use the retrieval capability of the toolkit and do not require the language modeling capability of it. Since our experiments were mainly targeted to provide insight into occurrence based statistics both at lexical level and syntactic level, we chose the same amount of data for both the experiments. As can be seen large corpus will results in a much broader coverage and aid in better analysis of the features. However, running parsers on a large scale to obtain accurate annotation data for a large scale of corpus is time and resource consuming. We therefore perform our experiment with 500K sentences from the English side of the Europarl corpus. A parsed version of the Europarl corpus was obtained from (Univ of Edinburgh). We then extracted the dependency tree structures from

the parsed corpus and extracted all possible headword chains from the same.

Therefore our data repository consisted of 500K English sentences to extract lexical occurrence based statistics from and 500K parsed and headword chains extracted sentences to obtain syntactic occurrence based statistics from.

### 7.2 Features

The features extracted were of two kinds. A set of lexical occurrence based statistics and a set of syntactic occurrence based statistics. As discussed in (SECTION 4) we consider lexical occurrence based statistics at 7 levels starting from bigram, trigram to eight grams. Similarly we consider syntactic occurrence based statistics at 7 levels, starting from bichain, trichain to eight chains. Although eight chains are quite sparse and most of the extracted statistics at eight chains are zero, primarily because the sentence from which we extract does not contain a headword chain of length 8, and even if it does, it is never seen in the corpus that we perform lookups on.

### 7.3 Correlation with Goodness of Translation

Automatic MT Evaluation is a major research problem and has been addressed by various research groups. The task is to compare automatically the goodness of the translation which is usually judged by the adequacy of the translation in conveying the meaning of its source and the fluency of the translation as per the target language. Automatic evaluation methods are still long way from making a highly accurate judgement, although metrics like BLEU (Papineni et al., 2001) and METEOR (Banerjee and Lavie, 2005) calculate scores that highly correlate with human output. In this experiment we do not compare our correlation with the scores of BLEU or METEOR.

The most important criterion for new features that aim to improve quality of MT outputs is to bear some sort of correlation with the quality of goodness of translation. In this experiment we try to compare the Spearman

Type	Sys1	Sys2	Sys3	Sys4
2gram	0.072	0.147	0.064	0.071
3gram	0.091	0.221	0.53	0.023
4gram	0.134	0.231	0.034	0.093
5gram	0.344	0.391	0.138	0.170
6gram	0.459	0.492	0.423	0.388
7gram	-	0.523	0.576	0.488

Table 1: Correlation of NGram occurrence statistics with Fluency

Type	Sys1	Sys2	Sys3	Sys4
2chain	0.085	0.196	0.132	0.060
3chain	0.016	0.152	0.103	0.071
4chain	0.253	0.450	0.244	0.205
5chain	0.502	0.510	0.439	0.477
6chain	0.519	-	0.569	0.538
7chain	-	-	-	-

Table 2: Correlation of Headword Chain occurrence statistics with Fluency

correlation of each of these extracted features with the actual human judgements of 'fluency' and 'adequacy'. We believe that a score obtained by some linear combination of these features will have even higher correlation with human judgements.

### 7.3.1 Fluency

A fluent sentence is one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker of English.

### 7.3.2 Adequacy

The judge is presented with the gold-standard translation and should evaluate how much of the meaning expressed in the gold-standard translation is also expressed in the output translation.

## 7.4 Syntax vs Lexical occurrence based features

In order to explore the discriminative power of each of the syntactic and lexical occurrence based features, we trained individual classifiers based on these features to classify hu-

Type	Sys1	Sys2	Sys3	Sys4
2gram	0.123	0.161	0.131	0.044
3gram	0.085	0.181	0.129	0.044
4gram	0.156	0.184	0.142	0.068
5gram	0.373	0.329	0.266	0.124
6gram	0.493	0.506	0.459	0.382
7gram	-	0.553	0.563	0.534

Table 3: Correlation of NGram occurrence statistics with Adequacy

Type	Sys1	Sys2	Sys3	Sys4
2chain	0.158	0.164	0.173	0.163
3chain	0.134	0.129	0.165	0.173
4chain	0.376	0.381	0.209	0.192
5chain	0.499	0.514	0.443	0.448
6chain	-	-	0.571	0.549
7chain	-	-	-	-

Table 4: Correlation of Headword Chain occurrence statistics with Adequacy

man translated and machine translated outputs. We selected the same MT 2003 data of 2 systems and 2 human outputs. Out of the 919 sentences for each of the system, we selected 400 sentences as training data and 200 sentences as testing data. Therefore the total training data consisted of 1600 sentences out of which 800 were human translated texts and 800 were from system translations. Similarly the test data had equal number of human and machine sentences. The results obtained are as shown below.

From the experiment, it can be seen that syntactic occurrence based statistics as features have greater discriminative powers when compared to pure lexical based occurrence features. We realize that better feature selection techniques and better scoring and feature combination functions will further improve the discriminative classification capabilities of the occurrence features in general.

## 8 Conclusion and Future Work

In this paper, we have proposed and investigated a new set of features called occurrence based statistics for machine translation. Occurrence based statistics are obtained by

Features	Accuracy	Prec	Recall
NGrams	52.2%	56%	44%
NChains	69%	72%	48%

Table 5: Classification Evaluation

looking up the existence of sequence of lexical units (ngrams) in huge human created language repositories without actually dealing with their probabilities. We have also experimented occurrence based statistics for certain syntactic units like the headword chains (nchains). Our experiments show that OBS correlate well with translation fluency and adequacy and are useful as discriminative features in classification task of identifying human translations and system translation from a set of sentences. We have calculated the OBS using a repository of 500K sentences from Europarl corpus, which is quite limiting. We would like to immediately scale this to a Giga word corpus and then to the entire Universal Digital Library of texts. We will also be experimenting with variations of Information Retrieval datastructures and algorithms that provide quick realtime retrieval for OBS. Our primary goal is using these features as a supplement to traditional language modeling to improving the decoding in a Statistical Machine Translation system.

## References

- Satanjeev Banerjee and Alon Lavie. 2005. ME-THEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 255–262, Morristown, NJ, USA. Association for Computational Linguistics.
- Philipp Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.
- Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.
- Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT-EMNLP '05*.
- D. Metzler, T. Strohman, H. Turtle, and W. Croft. 2004. Indri at trec 2004: Terabyte track.
- K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation.
- L. Shen, A. Sarkar, and F. Och. 2004. Discriminative reranking for machine translation.
- H. Yamada and Y. Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *IWPT2003*, pages 195–206.
- Ying Zhang and Stephan Vogel. 2005. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*, Budapest, Hungary, May. The European Association for Machine Translation.
- Ying Zhang and Stephan Vogel. 2006. Suffix array and its applications in empirical natural language processing. Technical Report CMU-LTI-06-010, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Dec.