# Machine Translation and Vernacular: Interpreting the Informal

Nora Tien

December 15, 2016

## 1 Introduction

Machine Translation (MT) is the automatic, or at least computer-aided, translation of text from one language to another. The history of MT dates back to the time of the first computers, and evolved from Rule-Based MT (RBMT) to Statistical MT (SMT). Both approaches contend with challenges stemming from linguistic, semantic, and contextual complexity, as well as, in the case of SMT, the necessity of robust training data in the source and target languages [3]. Operational difficulties such as evaluating and extending existing translation systems to new languages and domains further complicate the task of MT [2].

MT history had a promising beginning until the 1960s, when the Automatic Language Processing Advisory Committee criticized the low performance and relatively high cost of MT compared to human translation, leading to cuts in funding and research. However, as hardware advancements were made and globalization increased, interest in and support for MT has risen [2].

In this paper, I will summarize some of the linguistic and operational challenges that MT confronts. I will also discuss the word, phrase, and syntax based techniques that make up Rule-Based MT, as well as the semantics-oriented Interlingua approach. I will then cover, within the realm of linguistic-based models, Constraint, Knowledge, and, Principle Based, and Shake and Bake Models. Pertaining to Statistical Machine Translation, I will cover Alignment Models and their components, language models, bitext–a piece of text available in the source and target language with certifiably high quality translation–translation models, and decoders. I will also provide a brief survey of popular (statistical) models in the field: Log linear models, Example Based Models (EBMT) and Dialogue Based Models (DBMT). To conclude the history of MT, I will discuss manual and automatic evaluation techniques and error analysis [3]. Finally, I will describe the problem of informal language translation and my proposed solution, and provide a review of current efforts to address informal language in MT.

# 2 Linguistic Challenges

When translating one language into another, the structural, syntactic, and lexical differences must be taken into account.

## 2.1 Word Order

Word order, for example, varies by language–with English requiring adjectives come before nouns and Spanish requiring the opposite. In general, languages can be separated into three categories of word-order conventions. There exist Subject-Verb-Object languages such as German, French, English, and Mandarin, that typically have prepositions. There also exist Subject-Object-Verb languages such as Japanese and Hindi that typically have postpositions. Finally, there exist Verb-Subject-Object languages like Irish, Arabic, and Biblical Hebrew [2].

The difficulty posed by differing word order then, for MT, is in the necessary restructuring of sentences when going from the source to the target.

## 2.2 Tense and Pronouns

Marking of verbal tense is common in English, but often omitted in Mandarin–past, present, and future are often not explicitly stated. Spanish, too, differentiates between the past preterit and past imperfect (the simple past, something that occurred and terminated, and the continuous past)(1)(2) [2]. English also has more articles, as in (3), than, for example, Mandarin. Languages also differ in terms of what they omit. Pro-drop languages such as Japanese, Mandarin, and Spanish allow the omission of pronouns (4). However, even within these pro-drop languages, rate of omission–or referential density–varies. Cold languages consist of referentially sparse languages, necessitating more work to find the correct antecedents, while languages with more explicit pronouns are known as hot languages [2][3].

Thus, languages having divergent conventions regarding tense and pronouns, some expressing more information than others, complicates the word choice for MT. When going from a language with fewer tenses to a language with more, context is necessary to determine the appropriate target words. When moving from pro-drop to non-pro-drop, MT must recover the antecedents to generate the target sentence.

(1) yo hablaba mucho con él
'I used to talk to him a lot'
Past imperfect, ongoing in the past

(2) Una vez hablé con él
'I spoke with him once'
Past preterit, clear termination of event

(3) the, a, those

(4) quiero viajar, 'want to travel'
yo quiero viajar, 'I want to travel'
are both correct

## 2.3 Morphological Differences

Morphemes are the smallest units of a language that have meaning, and morphological difference manifests itself in two ways with respect to them; the number of morphemes per word and the way in which morphemes are altered or added together. Isolating languages like Vietnamese and Cantonese typically have one morpheme per word, while Polysynthetic languages have many morphemes per word, like Siberian Yupik, where a single word can be translated into an entire sentence in a more isolating language. The morphemes in agglutinative languages like Turkish and German have explicit boundaries, while, in fusion languages like Russian, affixes may combine morphemes [2].

In this way, morphological variety can hinder MT–where no morpheme boundaries exist, or where morphemes can change depending on how they are used and combined, the separation lines of the units of a sentence may be ambiguous. Furthermore, morphologically rich languages many have many different forms or conjugations of a single word, necessitating more information be stored per word for those languages in order to apply the appropriate translation rules.

## 2.4 Argument and Linking Structure

Differences also arise in argument structure, or argument to predicate linking. For example, where the head, in linguistics, is the word in a phrase that defines that phrase's syntactic type (noun, verb phrase) and the dependents are the words depending on the head in a phrase [11], Hungarian utilizes head marking (5)–the head reflecting the relationship between the head and its dependents. English, however, utilizes dependent marking (6), where the non-head bears the marking of the relationship. Linking may also vary in how abstract qualities of an event are delegated onto different words as well. Verb-framed languages use verbs to indicate direction of motion (7), while satellite-framed languages use satellites, or prepositions that are part of the verb phrase, to indicate direction (8). Another example of syntactic complexity lies in the linking of prepositional phrases to correct antecedents. For example, in the English sentence 'I want to see the actress with the daughter in the play', 'in the play' could apply to either actress or daughter. Depending on the language, this ambiguity can be preserved in translation, or may need to be resolved with aid from contextual clues [2].

Hungarian:

(5) az ember haz-a
the man house-his
'The man's house'
*house* is the head

(6) 'the dog's bone'
*bone* is the head

(7) Spanish to English:
acercarse, 'to come near'
salir, 'to go out'

(8) fly out
jump up
come in

Like word order differences, argument and linking Structure variation can require sentence restructuring, and a plethora of bilingual information detailing how to transform a particular language with specific linking rules into another with different rules.

## 2.5 Idiomatic Expressions

The appropriate part of speech of a target word may differ from the part of speech of the source word. For example 'is obligatory' may be best translated not to the Spanish 'es obligatorio', preserving the 'is-adjective' structure, but to the Spanish 'se obliga', the third-person reflexive conjugation of the verb 'obligarse', meaning 'to obligate onself' [2][3]. The target verb may change, as well, according to the idiomatic rules of expression (9). Idiomatic expressions also require the addition of arguments or prepositions in the target language when translating a source word or phrase(10) [2][3]..

(9) Hace frío
'it makes cold'
It's cold
*Spanish*

(10) Wo gei wo mama da dian hua
'I gave my mother a call on the phone'
I called my mother
*Mandarin pinyin*

Idiomatic expressions thus necessitate extensive, pairwise bilingual information to map source phrases to the correct target phrases, as idiomatic expressions typically defy rules and cannot be translated literally.

## 2.6 Gender

Additionally, some languages have many more gendered words than others–romance languages gendering most adjectives, while Mandarin has a single, gender neutral pronoun (in speech and transliterated text, although the third person pronoun character is gender specific) for the third person and virtually no gendered adjectives.

This illustrates the challenge, when between languages where one is more gendered than the other, the difficulty of recovering from context the gender of a word.

## 2.7 Lexical Complexity

Lexical differences consist of one word in the source language having multiple, corresponding words in the target language with orthogonal meanings. In English, this is created by homophones. A similar phenomenon occurs with polysemy, where a single source word may have multiple meanings, but the target language has different words for each of those

meanings. For example, the word 'to know' in English may be translated as 'saber' in Spanish and 'savoir' in French, meaning 'to understand' (e.g. a fact), or as the word 'conocer' in Spanish and connaître in French, meaning 'to be familiar with' (e.g. a person or place). Further, Mandarin and Japanese have different words for younger and older brothers and sisters, while English, and most romance languages do not distinguish. Lexical gaps exist as well, when one language may lack the word or phrase necessary to express a word in another language. For example, there is no equivalent in Japanese for the English word 'privacy', and there is no equivalent in English for the Mandarin word xio–the closest approximation being 'filial piety' [2][3].

Lexical differences and gaps thus require that we precisely define the meaning of a word in the source language–otherwise known as word sense disambiguation, a computational linguistics problem in its own right–to effectively translate it to the target language equivalent.

### 2.8 Cultural Idiosyncrasies

Cultural differences and stylistic tendencies also manifest themselves differently. For example, Mandarin names consist of words with meaning in everyday language, and can be translated literally or transliterated, which maintains the English convention of names that do not necessarily have everyday meaning (11)(12). Phrases or words succinct and elegant in one language may be unwieldy in another [3]. Thus, a thorough understanding of both languages is necessary. Machine translation can describe regularly occurring differences with general rules, but must handle idiosyncratic and lexical differences case-by-case. This latter category is known as translation divergence. Languages do have some commonality, however, as all appear to have words for talking about people and common activities like eating, drinking, sleeping. Furthermore, languages all seem to have parts of speech like nouns and verbs [3].

Idiosyncrasies of languages like these require choices to be made, on a language by language basis, that machines alone cannot make.

(11) Mandarin:
Yi Lan
'idea of many colors'

(12) English:
Michael, Theresa

## 3 Operational Challenges

Other difficulties arise in the realm of MT in terms of generalizing MT to new languages and contexts, wide ranges of writing styles, updating already built systems, achieving usability, and evaluating MT performance. To handle some of the semantic, contextual, and linguistic ambiguity, MT is often used over a limited domain with a lexicon that is a subset, say of size 8000, 12000 entries. Even so, a high level of linguistic knowledge is necessary, and is one of the main bottlenecks in Rule Based MT [2].

All of these typological characteristics contribute to the complexity and difficulty of translation. Some of these extreme kinds of complexity simply do not occur in common MT tasks, but nonetheless, automatically producing translations on par with human-translated results is still unsolved [2][3].

MT can be still be useful for non-literary translation, however, where simply communicating the meaning through a rough translation suffices, or where humans perform post-editing or the translation is limited to a specific domain with a subset vocabulary (e.g. restaurant recommendations, manuals). MT is also convenient for tasks necessitating a quick turnaround or with a high number of documents to be translated [3]. In these applications, coherent, albeit less faithful or elegant translations are still difficult, but achievable using the following MT approaches.

# 4 Fundamental Methods of Machine Translation

There exist a myriad of research paradigms that can shape MT. Three main themes: linguistically informed paradigms, linguistically agnostic systems, and systems using a combination of techniques from the previous two [2]. The following sections will cover the first two of these three approaches.

# 5 Linguistically Informed Models

## 5.1 Rules-Based MT

Broadly defined, Rule-Based MT (RBMT) is uses rules specific to each language pair to transform the source text to the target language. Thus, it is language-sensitive and requires copious amounts of overhead information–bilingual dictionaries along with pre-defined transformation rules (13). Within RBMT, there exist three main methods: Direct, Transfer, and Interlingua.

(13) Spanish: Noun Adj
English: Adj Noun

### 5.1.1 Direct

The direct approach consists of word-by-word translation of the source language document, employing an extensive bilingual dictionary and reordering as necessary after the word-by-word transformation. In (14), Direct RBMT might translate the sentence word by word to end up with the awkward English sentence, applying simply reordering rules (noun adj to adj noun) to restructure the sentence afterward. The bilingual dictionary would also correct the 'It does not please me' to 'I dislike'.

(14) No me gusta la poesía melancólica
'It does not please me the poetry melancholy'
I dislike melancholy poetry

### 5.1.2 Transfer

The transfer approach involves parsing the source text and using contrastive knowledge about the source and target language structural differences to generate the target language parse structure from the source language parse structure. For example, (14) might be parsed as

[[no me gusta]VP[la poesía melancólica]NP]S, where NP is a Noun Phrase, VP is a Verb Phrase, and S is a sentence. Without a bilingual dictionary, 'no me gusta' would not be mapped to 'I dislike', and this source parse structure might be transformed to NP VP for English target translation, and result in 'Melancholy poetry does not please me'–a coherent, albeit not necessarily fluent sentence in English.

Thus, the transfer model consists of analysis, transfer, and generation. More specifically these steps require the parsing of input text in the source language, the transformation of the input parse structure to the target structure, and the creation of output text based on the transformed input structure. A syntactic rule that might be used in the transfer model is word order. Lexical ambiguity and idioms might be handled with rules enforced by a bilingual dictionary, or by performing word sense disambiguation during the analysis phase.

In real applications of MT, these simple transfer rules do not suffice, however, and a mixture of direct and transfer models are used to provide the extensive cross-lingual lexical, syntactic, and semantic information necessary. In this case, the analysis stage involves morphological parsing, part-of-speech tagging, noun-phrase, prepositional phrase, and shallow dependency parsing. The transfer stage involves worse sense disambiguation, deciphering idioms, and grouping prepositions with their corresponding verbs. The synthesis stage then reorders, generates the target language morphemes, and applies the robust bilingual dictionary for lexical rules.

### 5.1.3 Interlingua

The interlingua approach extracts the meaning of the source language and generates the target language text from this abstract representation. Thus, without the need for language-to-language lexical, syntactic, and semantic rules and bilingual dictionaries, this approach is better suited for many-to-many translations. The resulting representation of meaning is the interlingua. The objective of this method is to assign the same representation, regardless of language, to all sentences with the "same" meaning. (14) might be represented as the interlingua depicted in Figure 1.

This method requires thorough semantic parsing, e.g. minimal recursion semantics, event-based representation, and breaking sentences and words into "atomic semantic primitives." The first connects events to their corresponding arguments using a minimal, fixed set of thematic roles. Thus, we must define abstract properties and represent non-event-based phenomena, like that of 'having' a color (e.g. the green witch). The extraction process, then, requires a great amount of work, making the analysis part of the interlingua method the most labor-intensive. Interlingua generalizes better however, and requires less rules. Even semantic ideas or concepts, however, can complicate the interlingua model when some languages require a higher degree of specificity, or have concepts that are not universal [3]. Otherwise, critics of the interlingua approach dislike the loss of syntactic, stylistic, and

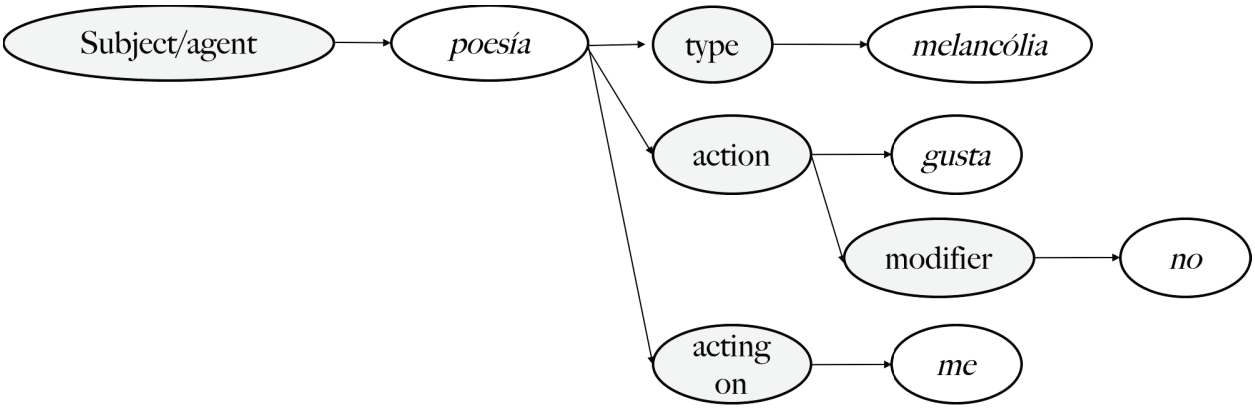


Figure 1: Interlingua representation of the sentence *No me gusta la poesía melancólica.*

emphatic information. However, for most MT applications, such knowledge is extraneous [2]. Figure 1 visualizes these three methods: as we move up the triangle, less and less transfer knowledge is needed.

# 6 Other Linguistic Models

## 6.1 Constraint Based

Constraint-Based MT (CBMT) applies constraints to a combination of lexical items. Specifically, CBMT maps betwee the source and target language structure, similar to the transfer approach, but using constraints encoded as lexical entries [2].

## 6.2 Knowledge Based

Knowledge-Based MT (KBMT) emphasizes access to extensive linguistic information–morphological, syntactic, and semantic. Additionally, KBMT requires a syntactic structure in the lexicon for all abstract events. This informational overhead makes KBMT expensive, as the abstract representation looks very little like the surface manifestation, and there exists no single, correct mapping [2]. In this way, KBMT reflects some of the characteristics and difficulties of the Interlingua approach.

## 6.3 Principle-Based

Principle-Based mirrors RBMT, but differs in that is utilizes a pared down set of principles rather than rules that describe the morphological, grammatical, and lexical aspects of the source and target language in a systemic, exhaustive manner [2].

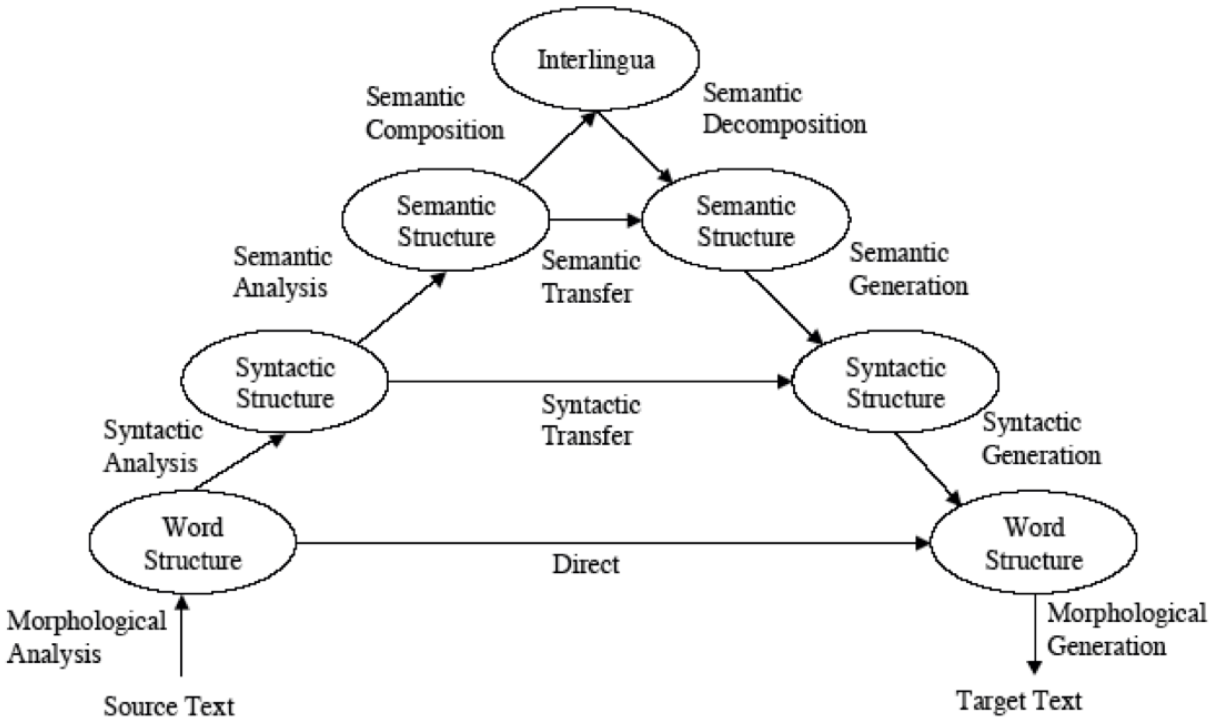


Figure 2: The Vauquois Triangle [3]

### 6.4 Shake and Bake

Shake and Bake (SBMT) employs conventional transfer rules to map lexical entries across languages, combining the resulting lexical items by using bilingual lexical dictionaries. SBMT extracts words from the source sentence parse structure, maps them to target words using the bilingual dictionary, and then rearranges these target words according to the target-language rules [2].

## 7 Linguistically Agnostic Models

### 7.1 Statistical MT

Ignoring linguistic information relies on sufficient training data (monolingual text and bitext). (Dorr). This is where Statistical MT comes into play. SMT emphasizes the ends, not the means, thereby precluding exact translation. Because of this, the constrains of translation must be relaxed. For SMT, we can represent translation as producing some target text that maximizes the objective function, incorporating both "faithfulness and fluency". SMT accomplishes this by creating probabilistic models of these two criteria and blending them to select the translation of highest-likelihood [3]. Specifically, the objective function comes from using Bayes' rule to transform the probability of the target text given the source text, $P(T|S)$, to the probability of the source text given the target text multiplied by the probability of that target text, $P(S|T)P(T)$. Thus, the translation probability achieves faithfulness

to the source meaning, and the lone probability of the target text, extracted from a language model, achieves fluency in the target language. In this way, combining these two models mitigates the overall error [6]. Specifically, SMT uses a model to translate. This model is defined by rules extracted from training corpora. Common architectures employed in the process of SMT are Finite State Transducers, Synchronous Context Free Grammars, and Hidden Markov Model (HMM) Alignment Models. All of these tools aid in decoding by shrinking the translation search space, and generate, rather than a lone output sequence, two output strings and the alignment between them [6]. Specifically, FSTs and SCFGs attempt to generate all of the possible permutations of translations from the source to the target. Thus, FSTs and SCFGs handle long distance re-orderings more adeptly than the HMM Alignment model, which tries to preserve locality of order in alignment. Additionally, SCFGs allows for including syntactic information by using productions for one language that incorporate syntax, and using productions that permit arbitrary permutations in the other language [6].

### 7.1.1 Finite State Transducers

Finite-State Transducers extend Finite State Automata. Formally, an FSA (S, L, D) is a set of states S, a set of labels L, and a set of transitions D. For each transition in D, there exists a subset S x S x L consisting of two matched states and a label that will be output, or consumed as the model travels from the first to the second state. Refer to Figure 3 for a simple example of an FST that consumes a string *a*b+* and outputs a string *(bba)*b*. An FST mirrors FSAs but with an added label set–each transition in D mapping to a label from both label sets. Thus, the FST reads an input string and writes an output string. In this way, x and y correspond to each other when a transition has label x from set L1 and label y from set L2. Either or both x and y may contain the empty token epsilon, signifying no change in the output string for the given transition. We can compose FSTs by taking the output of one and using it as the input to the other [6]. FSTs are limited in that they cannot reorder words long-distance without trying exponentially larger numbers of permutations.

### 7.1.2 Synchronous Context-Free Grammars

Synchronous Context-Free Grammars, or SCFG generalize Context-Free Grammars, and draw more from linguistic rules of syntax than FSTs. They are thus more capable of handling long-distance reordering without the high-cost of exhaustive permutations. As with FSTs and FSAs, SCFGs apply CGFs to the task of producing two output strings. Formally, an CFG may be defined as (N, T, D) with N nonterminal symbols, T terminal symbols, and D productions, such that D = N N union T*. In producing output, we begin with a root nonterminal symbol and recursively replace nonterminals via the production rules until we are left with only terminals. The SCGF grammar indicates two output strings for every production rule (indicated by co-indexed nonterminals). The co-indexing system describes the alignment between nonterminals of two languages. We can imagine SCGF's creating isomorphic trees, where the nonterminal nodes of the source and target are aligned, and the alignment between source and target words comes from the alignment of their parent nonterminal symbols. Refer to Figure 4 for an example of two parse trees, in German and

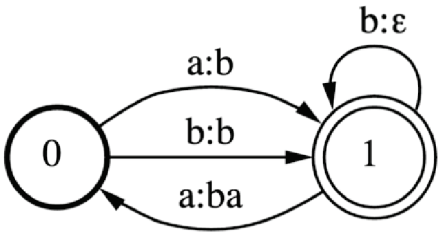


Figure 3: A simple FST [8]

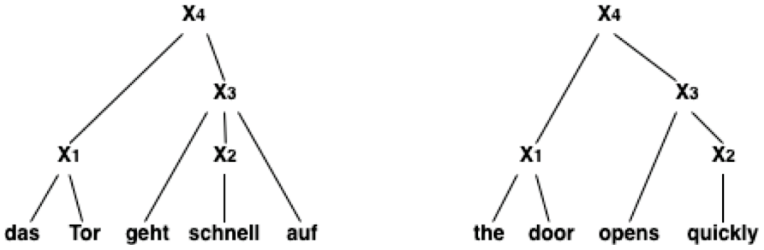


Figure 4: Parse Trees for German and English generated by a SCFG [9]

English, that might be output by a SCFG. Therefore, since subtrees are repeated recursively across the parse tree, SCGFs run in polynomial time (if utilizing dynamic programming) the exponential number of different reorderings [6].

### 7.1.3 HMM Alignment Model

These architectures aid in training the alignment models that MT relies on. The HMM alignment model currently dominates the industry, and relies on the paradigm that words close to each other in the source sentence are frequently aligned to words that are close to each other in the target sentence [6]. The HMM algorithm can preserve/prioritize this locality by training every alignment choice on the previous alignment choice [3].

Specifically, MT alignment models attempt to compute $P(T, A|S)$ (where T is the target word, A is the alignment, and S is the source word). The HMM approach transforms this probability into one based on the length $P(J|S_1^I)$, alignment $P(A_j|T_1^j - 1, a_1^j - 1, s_1^J)$, and lexicon $P(t_j|t_1^j - 1, a_1^j, s_1^I)$ probabilities using the chain rule. Further simplifying Markov principles are applied by assuming that the probability of a given alignment $a_j$ for a target word $t_j$ depends only on the aligned source word $s_{a_1}$ at position $a_j$. Lastly, we assume that we can estimate the length probability as $P(J|I)$ (where J is the length of the source sentence, and I is the length of the target sentence).

probability of target word $t_j$ depends only on the aligned source word $s_{a_j}$ at position $a_j$:
$P(a_j|t_1^j - 1, a_1^j - 1, s_1^I) = P(a_j|a_{j-1}, I)$

$P(t_j|t_1^j - 1, a_1^j, s_1^I) = P(t_j|s_{a_j},)$

Thus, the probability model for HMM alignment is:
$P(t_1^J, a_1^j|s_1^I) = P(J|I) \prod_{j=1}^J P(a_j|a_{j-1}, I)P(t_j|s_{a_j})$

To extract the total probability of the target sentence $P(t1^J|s1^I)$ we must sum over all

alignments.

$P(t_1^J|s_1^I) = P(J|I) \sum_A \prod_{j=1}^J P(a_j|a_{j-1}, I) P(t_j|s_{a_j})$

In this way, the previously aligned word informs the alignment, and we capture locality. More advanced HMM implementations condition alignments not on absolute word position but on relative distance between word positions, and add NULL source tokens to align with target words without mappings to source words, or incorporate the class of a word to informing the alignment [3].

### 7.1.4 Decoding

The next step in translation–after extracting alignments from the alignment model–is decoding. Decoding is essentially a search task. Given new input sentences, decoding attempts to solve the maximization problem: $s = argmax P(\hat{s}, d|t)$. $P(s, d|t)$ ranges over $S * xD * xT*$. T is fixed and the possible values of $(s, d)$ are bounded by the translation model, but there are still many $(s, d)$ values to compare. Thus, decoding attempts to search the space of all possible translations efficiently [6], typically using the Viterbi algorithm [3].

**FST and SCFG Decoding**

In FST Decoding, the search moves through a directed acyclic graph of states corresponding to partial or full translations generated from left to right in the target language word ordering. In SCGF decoding, the best-scoring tree that produces the input string using the source set of the grammar is found, and that tree is then output in the target language word order [6].

### 7.1.5 Log-linear models

Log-linear models incorporate both language and translation models to directly search for the target sentence with the best posterior probability: $S = armgax P(S|T)$. This is accomplished by defining P(S—T) with a set of features with parameters $\lambda_m$. The translation probability is then defined as:

$$P(S|T) = \frac{exp[\sum_{m=1}^{M} \lambda_m h_m(S,T)]}{\sum_{E^t} exp[\sum_{m=1}^{M} \lambda_m h|m(S^1,F)]}$$

In this way, the noisy channel model parameters–language and translation model–still dominate in terms of their influence on the model, but the log-linear model allows the inclusion of other features. For example, feature functions can be defined for a reverse translation model (target to source), word penalty, phrase penalty, and unknown word penalty. Often, however, log-linear models are trained explicitly to minimize evaluation metrics. This practice is known as Minimum Error Rate Training, or MERT [3].

Although these tools–HMM Alignment models and Log-linear models in particular–are industry favorites, other SMT paradigms exist, and provide diverse insight into the task of MT.

# 8 Other Linguistically Agnostic Models

## 8.1 Example Based

Example-Based MT (EBMT) requires a collection of source to target translations, searches these for most similar source sentence to a new source sentence. The target sentence for this previously seen source sentence is then reconfigured to generate a new target sentence for the translation (Dorr) Thus, this method is very dependent on bi/multilingual data base.

## 8.2 Dialogue-Based

Dialogue-Based MT (DBMT) is meant to be used in direct contact with the original author of the source text. DBMT asks the user for input to clarify and specify the source input text to improve the translation, thereby creating source text that it can handle, and omitting the features outside its expertise to be dealt with by the author (Dorr).

SMT is ill-equipped to manage contextual dependencies occurring far apart in the source sentence. EBMT fails when faced with structural complexity. Often, linguistic information is used to parse the input text and statistical/example methods are used to recover word dependencies and generate phrase mappings. SMT trigram models have been used for this type of lexical translation [2]

# 9 Evaluation of MT

Evaluation and error analysis of MT are equally significant–and challenging–areas of research. Currently, evaluation can be carried out by human judgement. Translations are mainly evaluated based on how well they communicate the source language meaning and how fluently they express that meaning in the target language. Further metrics are based on how well a given human task can be carried out using the machine-translation.

Otherwise, in SMT, evaluation can be conducted with hold-out data from the bitext; data not used to train the model. Another per-word metric is Levenshtein or edit-distance. This describes the number of insertions, deletions, and substitutions necessary to transform an output word to the correct word. This does not account for word-ordering, however. The commonly-cited Bilingual Evaluation Metric (BLEU) makes up for this particular feature of MT [6]. Specifically, BLEU rewards single word matching between output and reference strings, along with n-gram matches up to some maximum value of n. In this way, it favors local word order that reflects reference local word order. Although it is commonly used, it often performs poorly when compared to human judgement, or when comparing systems with very different architectures. Thus, it is most useful for evaluating changes made to a single system over time, or very similar systems [3].

# 10 Problem Statement and Motivation

Machine Translation alone has proven eminently useful as globalization increases and the demand for texts to be made available in a multitude of languages rises, especially with the

common task of information retrieval on the internet [3]. The relevance of translation of informal language is tied to the increase of informally expressed information available on the World Wide Web. Social media platforms, micro-blogging, and internet forums account for much of this unstructured text, and current translation engines are not trained to cope with the departure from formal language [4]. MT models are traditionally trained on such formal documents as government proceedings, with high quality human-translated copies available. Without these resources, MT models struggle with the noisy nature of vernacular text [3][4][7]. For example, translating such a sentence as "It was like, almost over, so I was like, Ima leave now" is ambiguous. The words 'like', and 'Ima' are colloquial, and, without training data or a bilingual dictionary entry mapping to correct target words, would confound the MT system.

Translation of informal language has some overlap with other problems in machine translation as well, such as translating under-resourced languages (lack of bitext), accurate domain-specific translations, generalizing a translation model to multiple languages, and translation of speech and dialogue (which is often informal) [6]. In particular, I am interested in tackling the problem of informal language MT with a combination of normalization and treating informal to formal language as a translation task in and of itself, along with the interlingua approach.

# 11 Methods and Techniques

I intend to use the Python Natural Language Tool Kit open source library for preprocessing and other NLP tasks, and the Python open source Machine Learning library scikit-learn for implementing SMT models (HMM). I also intend to use publicly available datasets for model experimentation. These tools and methods will change as a more precise solution to translating informal language becomes defined.

# 12 Related Work and Applications

## 12.1 Challenges Specific to MT of Informal Language

Some work in improving the quality of MT of informal language has been carried out in industry and academia. Although the approaches may vary, the challenges faced are consistent. In particular, informal text contains ambiguous usage of words, abnormal or rare morphological forms, phonetic substitutions, spelling and punctuation errors, abbreviations, contractions, letter dropping or repeating, and neologisms–words that have recently appeared and spread but that have yet to be incorporated into formal language [10]. Thus, the nature of informal language, as well as its extreme fluidity pose challenges to MT, and undermine the effectiveness of both Rule-Based and Statistical approaches.

## 12.2 Normalization of Informal Language

As Wang summarizes, current approaches attempt to normalize or "restore" informal language to formal language and translate out-of-vocabulary (OOV) words and abbreviations.

Normalization approaches usually involve word-by-word spell checking methods, and typically increased the BLEU score of the normalized language to the target language translation [10]. These methods often depend on easily discernible word boundaries, however, and therefore perform better on alphabetic, space-delimited (e.g. English, Spanish) rather than logographic (Mandarin, Japanese) languages. Additionally, the emphasis on normalization ignores word context and assumes that the bulk of informal language consists of typos and grammatical errors. Automatic-Speech Recognition approaches lend themselves to this problem as well, if we assume that much of informal language contains phonetic substitutions?making words closer to their phonetic representation than their conventional written form. Otherwise, transforming informal to formal language can be treated as a translation task. Thus, a phrase-based SMT model can be applied, but requires sufficient training data–bitext between informal sentences and their formal sentence translations[10]. [1] found this approach to be effective for normalizing Dialectical Arabic to Modern Standard Arabic, as Arabic is a morphologically rich language, and the differences between informal and formal can often be described in morphological terms. Specifically, when looking at user generated content (UGC), Arabic has many dialectic forms (Dialectal Arabic, DA), and users rarely communicate in Modern Standard, or formal Arabic (MSA). [1] trained a model on hand-corrected UGC texts, and then performed MSA to English translation. The baseline model, not using DA to MSA correction, had a Word Error Rate (WER)–based on Levenshtein or edit distance–of 40.46%. Incorporating DA to MSA correction, however, decreased the model's WER to 31.32%. This represents a nontrivial improvement, and illustrates the effectiveness of SMT applied to informal language, although sufficient training data remains necessary [1].

## 12.3 Out of Vocabulary Words and Crowdsourcing

The task of translating OOV words and abbreviations has likewise inspired a variety of approaches. OOV words arise when words occur in input sentences that lack translations in the phrase table produced during training. As bitext is scarce, infrequently updated, or gathered from different domains, this is a common challenge within MT. Current research attempts to transform OOV words into in-vocabulary (IV) words using morphological knowledge, which relies heavily on morphological knowledge of specific language pairs and information (thus not relevant to morphologically simple languages). It further assumes that OOV words are merely transliterations of IV words rather than accounting for neologisms. In other approaches, IV paraphrases can replace OOV words. This is accomplished by "pivoting" through other languages or extracting distributionally similar IV words from monolingual training data. This technique requires sufficient bitext for pivoting, and, in the case of paraphrasing, can find distributionally similar phrases that may be semantically orthogonal. This treatment of OOV words often decreases the WER. Nonetheless, these deficits, along with the consistent realization of a dearth of training corpora, have motivated some researchers to focus on novel ways to collect accurate training data rather than new techniques to improve informal language MT. Wang, in particular, proposes the idea of crowd-sourcing for extracting informal phrase translations across languages, rather than informal to formal restoration. He proposes doing so by creating a social-network hosted, collaboration-oriented flash card game that allows users to learn the language and modify the flash card translations between

languages, as well as ranking the quality of translations [10].

## 12.4 Pre-editing Rules

Expanding on the idea of informal to formal transformation, but combining this with crowd-sourcing, Gerlach proposes creating pre-editing rules, and crowd-sourcing the application and evaluation of these rules. In her work, these efforts proved useful in cutting down on post-editing time of translations of user-generated content in forums. In particular, pre-editing rules for informal French text were created with regular expressions, morphological rules, and POS tagging. They were then applied by both experts and users with only slight difference in performance. The raw text, when translated into English, achieved a BLEU score of 50.97%, while the pre-edited text received a BLEU score of 50.92%. However, the BLEU score was also shown to have a less than 20% correlation to human evaluation of the pre-edited translations. Thus, although pre-editing did not affect the BLEU scores of the subsequent translations, it did reduce post-editing time substantially, along with edit distance. However, as with the previously described normalization techniques, these pre-editing rules do not encompass all anomalies and manifestations of informal language. Furthermore, Gerlach notes that evaluation of clarifying rules was difficult, as improved readability is not easy to quantify. More significantly, the use of a single target language (English) limited the usefulness of the pre-editing rules, but was necessitated by the absence of comparably well-developed SMT systems for French to other target languages. Participation and behavior of users in real-life must also warrants further study, as conveying the importance of pre-editing rules is challenging [4].

## 12.5 CLIR

Cross lingual information retrieval (CLIR) endeavors to fill in the gaps of informal to formal translation in a different manner. In Pseudo-Relevance Feedback (PRF), a search query is used to find top n relevant documents (based on word similarity to search query). Significant terms from these top-n most relevant texts are then used to expand the search query and retrieve a new top-n set. In CLIR, query translation and document translation can be combined to minimize errors caused by informal source language queries. Both document and query translations are assumed to be error-prone, with high volume of mistranslated, or altogether untranslated terms. A new technique for CLIR involves evaluating intra-language and inter-language feedback on a per-query basis. Intra-language feedback functions by searching for translated documents with a translated query, and then supplementing the translated query with terms from the translated documents. Thus, it relies on high-quality query-translation and can mitigate poor quality of document translations. Inter-language feedback uses the original query to find source language documents, find the corresponding translated documents, and then use those translated documents' significant terms to supplement the translated query. In this way, it improves the results for poorly-translated queries. In their paper, [5] propose, for every query, to evaluate which approach is better and proceed accordingly [5], yielding improved model performance.

# 13 Conclusion

The vast range and complexity of linguistic challenges hinders the fluency of any automated translation, and human quality translations appear to be, at least for now, out of reach. MT can still benefit tasks that require coherence over elegance, or joint human-machine translation projects, however, proving its continued utility.

Machine Translation moved from RBMT to SMT, with implementations in industry often incorporating both. SMT, however, currently dominates state of the art systems. In fact, many believe MT to be solved, at least in reference to formal language, or language pairs for which sufficient bitext exists. Regarding informal language and its dearth of training data, however, SMT falls short.

Current approaches involve transforming informal language to formal language through automated or crowdsourced correction, or acquiring more data through novel crowdsourcing based techniques. Thus applying SMT or Direct or Transfer RBMT to the task of translating informal to formal language, monolingually. The Interlingua approach has been essentially discarded–even the formal definition of Interlingua is nebulous, and current implementations are opaque. In fact, the notion of Interlingua as a universal meaning may be itself disputable. However, I think it valuable to attempt to test out this approach and its validity. In the past, as better, more accessible methods were easily applicable, Interlingua was not viable. Informal language, however, represents the perfect opportunity to attempt the creation of an Interlingua translation system.

Immediate next steps for this project would include creating a toy dataset of informal language sentences with formal language counterparts in the source language (English), as well as formal and informal translations of these sentences in a target language. This toy dataset would allow me not only to run an Interlingua model on verifiably informal text, but also to evaluate the model's efficacy. To proceed, I would investigate different semantic parsing methods, and what, if any, open source systems exist that might offer some insight into implementing an Interlingua approach for informal text. Ultimately, I aim to determine whether or not Interlingua MT performs well enough to warrant its high cost, specifically when comparing this cost to the alternative of waiting for, or manually generating the necessary training data for more widely used SMT methods. If so, then the Interlingua approach would stand as a viable solution to the problem of informal language MT.

# References


[1] Haithem Afli, Walid Aransa, Pintu Lohar and Andy Way. April 2016. "From Arabic User-Generated Content to Machine Translation: Integrating Automatic Error Correction". url: http://www.computing.dcu.ie/ away/PUBS/2016/Haithem.pdf. Accessed 9/30/16.

[2] Bonnie J. Dorr, Pamela W. Jordan, John W. Benoit. December 1998. "A Survey of Current Paradigms in Machine Translation." url: http://www.umiacs.umd.edu/users/bonnie/Publications/newai98.pdf. Accessed 9/14/16.

[3] Daniel Jurafsky and James H. Martin. Draft of 2007. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. (Draft) Chapter 25 on Machine Translation. url: http://stp.lingfil.uu.se/ santinim/ml/2014/

[4] Johanna Gerlach. March 2015. "Improving Statistical Machine Translation of Informal Language: A Rule-based Pre-editing Approach for French Forums." Accessed 9/30/16.

[5] Chia-Jung Lee, W. Bruce Croft. 2014. "Cross-Language Pseudo-Relevance Feedback Techniques for Informal Text." url: https://pdfs.semanticscholar.org/8c10/f695eaadff166e11178a82b677690d868e00.pdf Accessed 9/30/16.

[6] Adam Lopez. 2008. "Statistical Machine Translation". ACM Comput. Surv., 40, 3, Article 8 (August 2008), 49 pages DOI = 10.1145/1380584.1380586 http://doi.acm.org/10.1145/1380584.1380586.

[7] Adam Lopez and Matt Post. 2013. "Beyond bitext: Five open problems in machine translation." url: http://cs.jhu.edu/ post/papers/lopez-post-bitext13.pdf. Accessed 9/14/16.

[8] Mehryar Mohri. 1997. ?Finite-State Transducers in Language and Speech Processing. ACL. http://www.cs.nyu.edu/ mohri/pub/cl1.pdf

[9] Moses Statistical Machine Translation System tutorial url: http://www.statmt.org/moses/?n=Moses.SyntaxTutorial

[10] Aobo Wang. September 2011. "Informal Language Crowd-sourcing and Translation."

[11] Arnold M. Zwicky. August 1984. "Heads". J. Linguistics., 21, 1-29.