THE NAMING OF THINGS AND THE CONFUSION OF TONGUES

FLORENCE REEDER^
MITRE Corporation
7515 Colshire Dr.
McLean VA 22102
freeder@mitre.org

ABSTRACT
If we accept language as an endogenous system, then we can start discussing representations for language processing which have a basis in endogenous systems. This paper is a start in the direction of showing how such a model might be constructed – drawing from across disciplines such as psychology, psycholinguistics and neuroscience. What we propose, and show the start of, is a model of lexical items which rather than being a list of words and features, is a system of evidence points. We outline the underlying technologies supporting this and describe what each will bring to the model. We believe this model will better support data analysis where language values are involved and show how it could work in cross-language information analysis. The first area of application involves named entities – people, places, associations – which are frequently necessary intelligence analysis data points, but which have confounded systems designed to automatically incorporate them.

INTRODUCTION

Machine Translation (MT) systems translate between human, natural languages. Weaver [1947] described the translation process as a kind of noisy channel decryption process: a French document is really an English document coded in French. For instance, the literal translation of **omelette de frommage** is **omelet with cheese**. However, the process of translating between languages is more than the substitution of one word for another. Consider the example, **escuela de derecho de Harvard**. Literally translated, it is **school of the right of Harvard** instead of the more proper **Harvard Law School**. The message can be found, but it takes more work on the part of the message's receiver.

Traditional MT systems have been developed following the Chomsky ideal [1957] of modeling language through successive layers of processing, each providing another level of information which isolates words from meaning. The levels, from least complex to most complex, are: phonemic, realizing words as sounds; morphological, capturing features of words; syntactic, assigning structure; semantic, attributing meaning; discourse, describing interaction rules; and pragmatic, concerning the mechanics of language in use. As processing moves toward pragmatic, the work to transfer between languages is reduced, while the understanding and generation processes increase in complexity.

The most complex, interlingua represents a pure translation process where the analysis yields a language-independent representation from which generation can be done. The components which address these different levels include: a bilingual lexicon (or list of words and information about those words); a

grammar for the source and target languages; a set of transfer rules which translate structures between source and target representations.

To demonstrate the difficulties inherent in translation and why it accordingly resists traditional processing, even traditional rationalist thought, consider the word **bank**. It can mean a financial institution, a side of a river, the act of counting on an event, or driving around the corner – all dependent on the context[1] in which it is used. Translating **bank** into foreign languages is not straight-forward, as other languages can have different words corresponding to the different concepts related to this word. Because of the many-to-many mapping of words, word for word substitution is insufficient, therefore, a lexical transfer system translates poorly. While each successive level of processing supplies more features for disambiguation, we are still left with the problem of finding just the right set of words in a new language to capture the same message as conveyed in the source language.

The necessary amount of target and source background knowledge has increased as has the preparation cost. Also, we still do not have sufficient representations for the transformation of language at the conceptual level. To say **I am banking on the horse coming in** implies a concept which literal translation may not be able to account for. In this case, building an interlingual representation is an appropriate choice. As described elsewhere [Reeder, 2000], the process of developing an interlingual representation has been like chasing the "Holy Grail". Human translation is considered a creative process of both interpretation and conveyance of a given message. This paper argues that it is because MT (and in fact language) is an endogenous process, the traditional methods of representing language information will necessarily be insufficient We must look to new processing ideas to accurately perform the translation task.

MT AS AN ENDOGENOUS PROCESS

Endogenous systems have been described as those which are perform operations which are incomputable, non-algorithmic and irreducible [Rosen, 2000]. These types of systems have been shown to be self-referential and logically tractable [Kercel et al., 2001]. We now discuss language as an endogenous process and one particular language problem, that of the mechanical translation between languages, as a problem suited for this category.

MT system development has relied on the Chomsky tradition of language processing levels. These levels are based on the notion that language, and therefore the automatic processing of it, is reducible. MT systems typically break a paragraph into sentences, sentences into phrases, phrases into syntactic parts, syntactic parts into words and words into meaning, often a logic-based representation. Each stage of processing can be accomplished by rule-based or statistical methods or a combination thereof that typically starts with an assumption of word independence. Yet, This is not a realistic view for language processing, or the MT systems derived from it. We will now show how MT systems rightly fit into the category of endogenous systems.

---

[1] Where context can be the surrounding text, the conditions in which the message is uttered and the states of mind of the speaker and hearer.

The human facility for language has been demonstrated to be an endogenous [Kercel, 1999] system by experts and even, grudgingly, by language practitioners. As Kercel et al. [2001] show:

> *Chomsky admits that the use of language is not explainable or understandable by this reductionist strategy. Consequently, he fears that it may be infeasible to study significant problems of language communication. His fear would be well founded if theoretical modeling were to stay limited to the reductionist strategy. Thus, even from a reductionist starting point, Chomsky allows the possibility that new intellectual tools might be needed, and might be feasible.*

If we accept that language itself is an endogenous system, then MT qualifies as endogenous. For instance, meanings can be compositional: **cheese fries** literally interpreted is either **fries with cheese on them** or **cheese is a substance that can be fried** or **fries composed of cheese**. At the same time, we have many wordings that are not compositional, defying a logical representation. The saying **He bought the farm** has a non-compositional, or idiomatic, meaning that is different from the sum of the parts. Yet, there is a reasonable and rational causality to idiomatic language usage. The phrase **thumb-rule** stems from the days where the **Rule of Thumb** determined the size of a stick with which a man could beat his wife.[2] So the entanglement of words is one reason for considering MT under the endogenous system model.

If MT can be viewed in the impredicative model, then we can apply aspects of endogenous modeling to build better MT systems[3] – such as a record of the past; prediction of the future; inferential elements; causal elements in our ontology; and anticipatory behavior. A record of the past can be gleaned from corpora as in statistical language models. To achieve the next level of capability, we require a new way of looking at computing translations because we are modeling a process that is impredicative. The question, then, becomes how to utilize the best of the traditional models? Or should we even try this? What are the pieces of this overall puzzle? Can we find partial solutions that are effective and efficient?

TYING IT ALL TOGETHER

The questions just presented aim to find ways to arrive at a model for an MT lexicon to support the MT process which reflects the endogenous paradigm. While we are still grounded in computation,[4] we believe that there are reasonable, implementable approaches to explore these questions. We look to data mining and data modeling, reasoning approaches for combining multiple evidence sources [e.g., Schum, 1994], traditional lexicon development, language learning and evaluation, and psycholinguistics (to include neurolinguistic programming). The end goal is a representation which supports the learning and combination of multiple pieces of evidence to contribute to the "meaning" of a concept for the purpose of translation between languages.

---

[2] No greater than the width of his thumb.

[3] Approximating endogenous systems until actual capabilities are available.

[4] After all, as engineers, we do have to build **something.**

Accepting MT as an endogenous problem, we look for ways in which we can design intelligent systems which perform more successful MT. We need an approach which allows us to combine results from different computational representations of language and suggest an evidential lexicon as one possible basis. An evidential lexicon starts with information from multiple sources: dictionaries, corpora, analyses of language in use, psycholinguistics, etc., and results in a lexicon which facilitates intelligent word selection for translation.

Because many biological systems are endogenous, we look to psycholinguistics for ideas in modeling the impredicative nature of MT. Psycholinguistic models suggest why we select the words and grammatical constructs we use. Language is one of our primary means of acquiring information – through talk, through reading, through words. Language is not the form of the representation, but words and the combination of these words with contextual clues are part of our knowledge structure. The notion that we remember a gist of what was said in place of the exact words means that there is something else going on. Professional translators show this often when they translate sentences according to general meaning instead of for exact words. In fact, it has been shown that there is a wide degree of variability even in translating "factual" reports [Farwell & Helmreich, 1996]. Another study [Al-Onaizan, et al., 2000] reports that translation can be a process of picking the words you know, guessing a context and applying that context to unseen words. There is reason to suspect that the translation process is different than the sum of the parts.

What we propose is a framework which supports the integration of multiple evidence points to contribute to word (or possibly phrase) meaning. In this framework, translation selection depends on evidence for/against particular meaning. The evidence points will come from: dictionaries, which represent a kind of jurisprudence about the meaning and translations of words; learned values from corpora, such as mutual information measures from information retrieval; values from ontologies, both human and automatically constructed; and neurolinguistic programming inspired preferences. Each of these sources would contribute weighted values for word-sense disambiguation and translation preference selection. Initially, a Bayesian representation would support this, although alternative representations, combinations and computations must follow from the impredicative nature of language.

A SPECIAL CASE – NAMES

Acknowledging the inherent difficulties in looking at only part of a function when the whole is greater than the sum of the parts, we look at a specific kind of entry for the lexicon structure described here. *Named Entities* are the set of proper names and numerical expressions such as times, dates, monetary expressions, or percentages. Since these carry critical content, content that is useful for summarization or topic identification, handling their translation well represents a necessary area for MT system advancement. Proper translation of named entities must ensure that they are a) not translated when proper names; b) rendered idiomatically rather than literally, observing standard naming conventions in the target language; and c) rendered in a format which is usable other processing. We describe each of these.

Translating a name correctly, i.e., **Helmut Kohl,** means not translating it, but rendering it in a form usable by English NLP software. **Kohl** is the German word for **cabbage** – a system could amuse, but not be helpful, returning **Helmut Cabbage**. The first goal of translation, then, is to ensure that proper names are not translated as common nouns, but that certain titles and parts of names are.

With organizations, a different strategy is necessary. As noted earlier, while **Escuela de Derecho de Harvard** is properly translated as **Harvard Law School**, it is literally **Harvard School of the Right**. The idiomatic translation is preferred for institution names. In this family of challenges is acronym handling. Some acronyms are translated and others not: a literal translation of the abbreviation for the former Soviet Union could be rendered as **SSSR**, whereas it is normally rendered by translating the expression into English (Union of the Soviet Socialist Republics) and then reducing the initials to **USSR**. On the other hand, the Basque separatist organization is generally referred to as **ETA**, based on the abbreviation of the not translated Basque title.

The transformation of names into something legible for the target language speaker must be handled in languages that use characters that have no English equivalent. For instance, some people will render the name for the former Soviet Union as **CCCP** (a look-alike representation of the Cyrillic characters). Rendering can involve translation of diacritics such as German umlauts or French accents, or it can involve other alphabets (Cyrillic, Arabic, Thai) or other writing systems (Chinese, Japanese). Assuming the name was not to be translated, the translation engine leaves it in Cyrillic characters instead of an English transliteration (or transcription). This means that a reader of the translation might miss **Gorbachev**. As Knight and Graehl [1998] point out, even uniform transliteration does not guarantee success in name recognition. The large number of transliterations for **Khadafy** is an example.

A SAMPLE IMPLEMENTATION

The field of translation of "named entities" lends itself well to an evidential representation. Because many translation conventions are just that, conventions, a system capable of learning the features and functions that determine the conventions represents a significant gain over other MT approaches. We present the beginnings of our testing of this theory. We started with 100 documents [see Reeder et al, 2001, for a more complete description] which were originally written in Spanish and then translated by two professional translators into English. This serves as our test corpus for exploring the named entity problem.

We took one reference translation from the corpus and manually annotated it for named entities. We then wrote matching software which would search for matches between the two parallel articles (REFERENCE and EXPERT). Given the list of names extracted from an article, the software found the names in the corresponding translation. In this way, we get a picture of the kinds of features which contribute to the translation of names and the weights of each.

Our testing has yielded some interesting evidence points. First, dictionaries do not typically contain proper nouns, instead giving common noun definitions. Therefore, we sought another means of establishing a baseline or jurisprudence that dictionaries can give. To do this, we used multiple human translations. The baseline for matching between human translations of named entities was only

about 80% for exact match. This means that our next evidence points come from a series of relaxations on the mismatch reasons. Other features that must be measured to determine if a name phrase is a proper translation of another include: diacritics (accent marks); capitalization; and numeric values. While the inclusion of these measures improves matching to 90%, this leaves a number of features for which we are still establishing weights, including: morphology (such as **Peruvian** versus **of Peru**); stop-words (**of** versus **for**) and synonyms.

We envision implementing evidence first as a Bayesian network, although since we have just completed the feature identification for evidence points, we may adjust this. Additionally, we look to continuing endogenous system work for a more appropriate reasoning framework.

CONCLUSIONS

Some domain specific terms are specialized names – such as chemical names. But what of the chemical precursors? Glass vessel is composed of common nouns and only has a technical flavor to it when applied in the chemical or biological domains. In addition, we are confronted by the widespread use of loan-words and adapted loan-words in foreign languages – particularly due to the fact that the recognized language of science is English. As this is preliminary work, there are many important challenges ahead. Some of the challenges we foresee are in marrying information from multiple data sources, finding the supporting technologies to mine data sources; handling less-commonly taught languages where a paucity of information exists.

REFERENCES

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D., Yamada, K. 2000. Translating with Scarce Resources. In *Proceedings American Association for Artificial Intelligence Conference, AAAI-2000.*

Chomsky, N., 1957, *Syntactic Structures*, Mouton Publishers, The Hague.

Helmreich, S. & Farwell, D. 1996. Translation Differences and Pragmatics-Based MT. In *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96).* Montreal, Canada.

Kercel, S., 1999, Embedded Intelligence: Bizarre Systems for the Steel Industry, Presentation to AISE – Sensors Meeting.

Kercel, S., Brown-VanHoozer, S. A., VanHoozer, W.R.. (in press) The Entangled Future of Foreign Language Learning. In The Future of Foreign Language Education in the United States, ed. Terry A. Osborn. Contemporary Language Education Series. Westport, CT: Bergin & Garvey.

Knight, K. & J. Graehl. 1998. Machine Transliteration. *Computational Linguistics, 24(4),* 598–612.

Reeder, F. 2000. ISA or Not ISA: The Interlingual Dilemma for Machine Translation. In *Proceedings of ANNIE-2000, Bizarre Systems Track.* ASME Press.

Reeder, F., Miller, K., Doyon, J., White, J. 2001. The Naming of Things and the Confusion of Tongues: An MT Metric. *Proceedings of Workshop on MT Evaluation, MT-Summit 2001.*

Rosen, R., 2000., *Essays on Life Itself.* Columbia University Press, New York.

Schum, D. A., 1994, *Evidential Foundations of Probabilistic Reasoning.*, John Wiley & Sons, Inc.

Weaver, W., 1947, Letter to Norbert Wiener, March 4, 1947, (Rockefeller Foundation Archives).

---

^ The views expressed in this paper are those of the author and do not reflect the policy of the MITRE Corporation.