Proactive Learning for Building Machine Translation Systems for Minority Languages

Vamshi Ambati
vamshi@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Jaime Carbonell
jgc@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

Building machine translation (MT) for many minority languages in the world is a serious challenge. For many minor languages there is little machine readable text, few knowledgeable linguists, and little money available for MT development. For these reasons, it becomes very important for an MT system to make best use of its resources, both labeled and unlabeled, in building a quality system. In this paper we argue that traditional active learning setup may not be the right fit for seeking annotations required for building a Syntax Based MT system for minority languages. We posit that a relatively new variant of active learning, Proactive Learning, is more suitable for this task.

1 Introduction

Speakers of minority languages could benefit from fluent machine translation (MT) between their native tongue and the dominant language of their region. But scarcity in capital and know-how has largely restricted machine translation to the dominant languages of first world nations. To lower the barriers surrounding MT system creation, we must reduce the time and resources needed to develop MT for new language pairs. Syntax based MT has proven to be a good choice for minority language scenario (Lavie et al., 2003). While the amount of parallel data required to build such systems is orders of magnitude smaller than corresponding phrase based statistical systems (Koehn et al., 2003), the variety of linguistic annotation required is greater. Syntax based MT systems require lexicons that provide coverage for the target translations, synchronous grammar rules that define the divergences in word-order across the language-pair. In case of minority languages one can only expect to find meagre amount of such data, if any. Building such resources effectively, within a constrained budget, and deploying an MT system is the need of the day.

We first consider 'Active Learning' (AL) as a framework for building annotated data for the task of MT. However, AL relies on unrealistic assumptions related to the annotation tasks. For instance, AL assumes there is a unique omniscient oracle. In MT, it is possible and more general to have multiple sources of information with differing reliabilities or areas of expertise. A literate bilingual speaker with no extra training can produce translations for word, phrase or sentences and even align them. But it requires a trained linguist to produce syntactic parse trees. AL also assumes that the single oracle is perfect, always providing a correct answer when requested. In reality, an oracle (human or machine) may be incorrect (fallible) with a probability that should be a function of the difficulty of the question. There is also no notion of cost associated with the annotation task, that varies across the input space. But in MT, it is easy to see that length of a sentence and cost of translation are superlinear. Also not all annotation tasks for MT have the same level of difficulty or cost. For example, it is relatively cheap to ask a bilingual speaker whether a word, phrase or sentence was correctly translated by the system, but a bit more expensive to ask for a correction. Assumptions like these render active learning unsuit-

Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing, pages 58–61,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

able for our task at hand which is building an MT system for languages with limited resources. We make the case for “Proactive Learning” (Donmez and Carbonell, 2008) as a solution for this scenario.

In the rest of the paper, we discuss syntax based MT approach in Section 2. In Section 3 we first discuss active learning approaches for MT and detail the characteristics of MT for minority languages problem that render traditional active learning unsuitable for practical purposes. In Section 4 we discuss proactive learning as a potential solution for the current problem. We conclude with some challenges that still remain in applying proactive learning for MT.

2 Syntax Based Machine Translation

In recent years, corpus based approaches to machine translation have become predominant, with Phrase Based Statistical Machine Translation (PB-SMT) (Koehn et al., 2003) being the most actively progressing area. Recent research in syntax based machine translation (Yamada and Knight, 2001; Chiang, 2005) incorporates syntactic information to ameliorate the reordering problem faced by PB-SMT approaches. While traditional approaches to syntax based MT were dependent on availability of manual grammar, more recent approaches operate within the resources of PB-SMT and induce hierarchical or linguistic grammars from existing phrasal units, to provide better generality and structure for reordering (Yamada and Knight, 2001; Chiang, 2005; Wu, 1997).

2.1 Resources for Syntax MT

Syntax based approaches to MT seek to leverage the structure of natural language to automatically induce MT systems. Depending upon the MT system and the paradigm, the resource requirements may vary and could also include modules such as morphological analyzers, sense disambiguation modules, generators etc. A detailed discussion of the comprehensive pipeline, may be out of the scope of this paper, more so because such resources can not be expected in a low-resource language scenario. We only focus on the quintessential set of modules for MT pipeline - data acquisition, word-alignment, syntactic analysis etc. The resources can broadly be categorized as ‘monolingual’ vs ‘bilingual’ depending upon whether it requires knowledge in one language or both languages for annotation. A sample of the different kinds of data and annotation that is expected by an MT system is shown below. Each of the additional information can be seen as extra annotations for the ‘Source’ sentence. The language of target in the example is ‘Hindi’.

- **Source:** John ate an apple
- **Target:** John ne ek seb khaya
- **Alignment:** (1,1),(2,5),(3,3),(4,4)
- **SourceParse:** (S (NP (NNP John)) (VP (VBD ate) (NP (DT an) (NN apple))))
- **Lexicon:** (seb → apple),(ate → khaya)
- **Grammar:** VP: V NP → NP V

3 Active Learning for MT

Modern syntax based MT rides on the success of both Statistical Machine Translation and Statistical Parsing. Active learning has been applied to Statistical Parsing (Hwa, 2004; Baldridge and Osborne, 2003) to improve sample selection for manual annotation. In case of MT, active learning has remained largely unexplored. Some attempts include training multiple statistical MT systems on varying amounts of data, and exploring a committee based selection for re-ranking the data to be translated and included for re-training. But this does not apply to training in a low-resource scenario where data is scarce.

In the rest of the section we discuss the different scenarios that arise in gathering of annotation for MT under a traditional ‘active learning’ setup and discuss the characteristics of the task that render it difficult.

3.1 Multiple Oracles

For each of the sub-tasks of annotation, in reality we have multiple sources of information or multiple oracles. We can elicit translations for building a parallel corpus from bilingual speakers who speak both the languages with certain accuracy or from a linguist who is well educated in the formal sense of the languages. With the success of collaborative sites like Amazon’s ‘Mechanical Turk’ [1], one

[1] http://www.mturk.com/

can provide the task of annotation to multiple oracles on the internet (Snow et al., 2008). The task of word alignment can be posed in a similar fashion too. More interestingly, there are statistical tools like GIZA [2] that take as input un-annotated parallel data and propose automatic correspondences between words in the language-pair, giving scope to 'machine oracles'.

3.2 Varying Quality and Reliability

Oracles also vary on the correctness of the answers they provide (quality) as well as their availability (robustness) to answer. One typical distinction is 'human oracles' vs 'machine oracles'. Human oracle produce higher quality annotations when compared to a machine oracle. We would prefer a tree bank of parse trees that were manually created over automatically generated tree banks. Similar is the case with word-alignment and other tasks of translation. Some oracles are 'reluctant' to produce an output, for example parsers tend to break on really long sentences, but when they produce an output we can associate some confidence with it about the quality. One can expect a human oracle to produce parse trees for long sentences, but the quality could be questionable.

3.3 Non-uniform costs

Each of the annotation tasks has a non-uniform cost associated with it, the distribution of which is dependent upon the difficulty over the input space. Clearly, length of the sentence is a good indicator of the cost. It takes much longer to translate a sentence of 100 words than to translate one with 10 words. It takes at least twice as long to create word-alignment correspondences for a sentence-pair with 40 tokens than a pair with 20 tokens. Similarly, a human takes much longer to manually create parse tree for a long sentence than a short sentence.

It is also the case that not all oracles have the same non-uniform cost distribution over the input space. Some oracles are more expensive than the others. For example a practicing linguist's time is perhaps costlier than that of an undergraduate who is a bilingual speaker. As noticed above, this may reflect upon the quality of annotation for the task, but sometimes a tradeoff to make is cost vs quality. We can not afford to introduce a grammar rule of low-quality into the system, but can possibly do away with an incorrect word-correspondence link.

4 Proactive Learning

Proactive learning (Donmez and Carbonell, 2008) is a generalization of active learning designed to relax unrealistic assumptions and thereby reach practical applications. Active learning seeks to select the most informative unlabeled instances and ask an omniscient oracle for their labels, so as to retrain the learning algorithm maximizing accuracy. However, the oracle is assumed to be infallible (never wrong), indefatigable (always answers), individual (only one oracle), and insensitive to costs (always free or always charges the same). Proactive learning relaxes all these four assumptions, relying on a decision-theoretic approach to jointly select the optimal oracle and instance, by casting the problem as a utility optimization problem subject to a budget constraint.

$$\begin{aligned} maximize\ E[V(S)]\ subject\ to\ B \\ max_{S \in UL} E[V(S)] - \lambda(\sum_k t_k * C_k) s.t \\ \sum_k t_k * C_k = B \end{aligned}$$

The above equation can be interpreted as maximizing the expected value of labeling the input set S under the budget constraint B. The subscript k denotes the oracle from which the answer was elicited under a cost function C. A greedy approximation of the above results in the equation 1, where $E_k[V(x)]$ is the expected value of information of the example x corresponding to oracle k. One can design interesting functions that calculate $V(x)$ in case of MT. For example, selecting short sentences with an unresolved linguistic issue could maximize the utility of the data at a low cost.

$$(x*, k*) = argmax_{x \in U} E_k[V(x)]\ subject\ to\ B \quad (1)$$

We now turn to how proactive learning framework helps solve the issues raised for active learning in MT in section 3. We can address the issue of multiple oracles where one oracle is fallible or reluctant to answer, by factoring into Equation 2 its probability

[2] http://www.fjoch.com/GIZA++.html

function for returning an answer. The score returned by such a factoring can be called the utility associated with that input for a particular oracle. We call this $U(x, k)$. A similar factorization can be done in order to address the issue of oracles that are fallible.

$$\begin{aligned} U(x,k) &= P(ans|x,k) * V(x) - C_k \\ (x*, k*) &= argmax_{x \in U}\, U(x,k) \end{aligned}$$

Since we do not have the $P(ans/x, k)$ distribution information for each oracle, proactive learning proposes to discover this in a discovery phase under some allocated budget B_d. Once we have an estimate from the discovery phase, the rest of the labeling proceeds according to the optimization function. For more details of the algorithms refer (Donmez and Carbonell, 2008). Finally, we can also relax the assumption of uniform cost per annotation, but replacing the C_k term in the above equations with a $C_{non-unif_k}$ function denoting the non-uniform cost function associated with the oracle.

5 Future Challenges

While proactive learning is a good framework for building MT systems for minority languages, there are however a few issues that still remain that need careful attention.

Joint Utility: In a complex system like MT where different models combine forces to produce the translation we have a situation where we need to optimize not only for an input and the oracle, but also the kind of annotation we would like to elicit. For example given a particular translation model, we do not know if the most optimal thing at a given point is to seek more word-alignment annotation from a particular 'alignment oracle' or seek parse annotation from a 'parsing oracle'.

Machine oracles vs Human oracles: The assumption with an oracle is that the knowledge and expertise of the oracle does not change over the course of annotation. We do not assume that the oracle learns over time and hence the speed of annotation or perhaps the accuracy of annotation increases. This is however very common with 'machine oracles'. For example, an oracle that suggests automatic alignment of data using statistical concordances may initially be unreliable due to the less amount of data it is trained on, but as it receives more data, the estimates get better and so the system gets more reliable.

Evaluation: Performance of underlying system is typically done by well understood metrics like precision/recall. However, evaluation of MT output is quite subjective and automatic evaluation metrics may be too coarse to distinguish the nuances of translation. This becomes quite important in an online active learning setup, where we add annotated data incrementally, and the immediately trained translation models are not sufficient to make a difference in the scores of the evaluation metric.

References

Jason Baldridge and Miles Osborne. 2003. Active learning for hpsg parse selection. In *Proc. of the HLT-NAACL 2003*, pages 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. 43rd ACL*, pages 263–270, Morristown, NJ, USA. Association for Computational Linguistics.

Pinar Donmez and Jaime G. Carbonell. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *CIKM '08*, pages 619–628, New York, NY, USA. ACM.

Rebecca Hwa. 2004. Sample selection for statistical parsing. *Comput. Linguist.*, 30(3):253–276.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the HLT/NAACL*, Edomonton, Canada.

Alon Lavie, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjós, Rachel Reynolds, Jaime Carbonell, and Richard Cohen. 2003. Experiments with a hindi-to-english transfer-based mt system under a miserly data scenario. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):143–163.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the EMNLP 2008*, pages 254–263, Honolulu, Hawaii, October.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proc. of ACL '01*, pages 523–530, Morristown, NJ, USA. Association for Computational Linguistics.