NLG is STILL Relevant to MT

Nizar Habash
Center for Computational Learning Systems
Columbia University
New York, NY, USA
habash@cs.columbia.edu

In the model of the machine translation (MT) pyramid, natural language generation (NLG) had a prominent role and was considered an essential part of the MT process for interlingual and transfer approaches. But with the rise of shallow statistical MT (SMT) approaches, the role of NLG has disappeared almost completely. In this abstract, I take the position that NLG is also relevant to SMT and SMT research. I start with a discussion of similarities and differences among components of NLG and SMT. I follow this with a discussion of some ideas of how NLG can contribute to future research in SMT.

Components in MT and NLG can be classified in terms of two orthogonal dimensions: depth and modeling approach. Depth refers to the level of representation the components operate on: shallow surface words in full inflected form or deeper linguistic representations such as morphological, syntactic and semantic annotations. In principle, there is no particular dependence among the different depths from a modeling point of view. For example, syntactic models can be done over inflected words, their morphological abstractions, or just parts-of-speech. The degree of trade-off between sparsity and simplicity differs for different languages with varying degrees of morphological complexity.

In terms of modeling approach, components can be built using manual human "learned" rules (HL) or machine learned (ML) rules. There are many possible instances in between these two extremes that can include different degrees of manual human interference in pure machine learning: restricting the machine to a specific space of linguistic rules or more abstractly to a formal space of allowable linguistic rules.

These two dimensions are orthogonal. HL components tend to be deep as in traditional "symbolic/rule-based" MT and NLG, but can also be shallow as in templates in NLG or word-based MT. ML components tend to be shallow as in phrase tables in MT and n-gram language models for MT and NLG (Brown and Frederking, 1995; Langkilde and Knight, 1998). But much research has been going on in using deeper representations in MT (Collins et al., 2005) and language modeling (Bangalore and Rambow, 2000).

All positions on both dimensions come with disadvantages – there is no obvious "sweet spot" on this two-dimensional continuum. For instance, both linguistic and surface models are prone to hallucinating/over-generating constructions that are suboptimal or flat-out wrong. These can be a result of ML misalignment or HL over-abstraction. Similarly, HL may be more concise and less redundant compared to ML, but it tends to miss a lot of "less obvious" cases. Additionally, both are prone to coverage limitations, whether it be the domain and genre of the corpus or the focus of the linguist producing the data to build the model.

It is possible to define a common framework for MT and NLG in which every component of MT or NLG can be categorized as either (1) meaning-preserving transformation, whether one-to-one or one-to-many (or even many-to-many) or (2) hypothesis reranking using features from any part of the system. In the context of meaning-preserving transformations, sharable resources between NLG and MT have to be monolingual. These include resources that map one language into itself such as categorial variation (Habash and Dorr, 2003) or WordNet-based synset expansion (Fellbaum, 1997); and resources that map from a deep representation to a shallower one such as morphological generators (Habash, 2004) or interlingual lexicons (Dorr, 1993). Components in the context of reranking can be shared by both MT and NLG, as in n-gram language models or syntax-based language models (Brown and Frederking, 1995; Langkilde and Knight, 1998; Bangalore and Rambow, 2000).

Within the framework described above, the main difference between end-to-end MT and NLG is their input representation: MT expects shallow surface words in a different (source) language whereas NLG expects typically a deep representation that is in principle language-independent. In the degenerate case of considering source-language shallow words as the input representation for NLG, NLG and MT would be equivalent. Therefore any integration of components distinctly from NLG into SMT must necessarily use deeper representations. This entails the presence of generation's dual process, analysis, to produce the representation on which NLG components will operate. This is an added cost for shallow SMT.

For shared components, such as language models, improvements done in the NLG community can transfer easily to SMT, obviously. However, deeper integration can be more involved. NLG components can be used to perform expansions on the target language side that fill gaps in SMT models. This is very helpful particularly when translating into morphologically rich languages such as Arabic (Habash, 2004) or Turkish (Oflazer, 1993). For these languages, the cost of modeling the complex but regular morphology is cheaper than acquiring more data. Expansions of the SMT hypothesis space using NLG components operating at deeper levels such as semantics are also possible (Dorr and Habash, 2002). Here, NLG components (operating on the target language through analysis and back generation) extend the SMT target-language search space with alternative paraphrases that are included in the reranking.

Although NLG is typically expected to be a later component producing the target language, this is not a necessity. For instance, models of syntactic ordering can be in principle separated from surface realization, which can be handled using shallow phrase tables. Here, unlexicalized syntactic ordering models from NLG can be used to order syntactic structures with source-language words before translating them using phrase-table entries. Also, linguistic expansions do not have to be limited to the target language: analyzing and back-generating the source language can provide alternative morphological forms and paraphrases that can increase the likelihood of matching against the SMT phrase table.

As demonstrated with the examples above, there is a large space of possibilities for integrating components and ideas from NLG into SMT.

References

S. Bangalore and O. Rambow. 2000. Exploiting a Probabilistic Hierarchical Model for Generation. In *International Conference on Computational Linguistics (COLING 2000)*, Saarbrucken, Germany.

R. Brown and R. Frederking. 1995. Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 221–239, Leuven, Belgium.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan.

Bonnie J. Dorr and Nizar Habash. 2002. Interlingua Approximation: A Generation-Heavy Approach. In *Workshop on Interlingua Reliability, Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, CA.

Bonnie J. Dorr. 1993. *Machine Translation: A View from the Lexicon.* The MIT Press, Cambridge, MA.

Christiane Fellbaum. 1997. *WordNet: An Electronic Lexical Database.* MITPress, Cambridge, MA.

Nizar Habash and Bonnie Dorr. 2003. CatVar: A Database of Categorial Variations for English. In *MT Summit*, pages 471–474, New Orleans, LA.

Nizar Habash. 2004. Large Scale Lexeme Based Arabic Morphological Generation. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN-04)*, pages 271–276. Fez, Morocco.

Irene Langkilde and Kevin Knight. 1998. Generation that Exploits Corpus-Based Statistical Knowledge. In *ACL/COLING 98, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (joint with the 17th International Conference on Computational Linguistics)*, pages 704–710, Montreal, Canada.

Kemal Oflazer. 1993. Two-level Description of Turkish Morphology. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, Utrecht, The Netherlands.