English-to-Mandarin Speech Translation with Head Transducers

Hiyan Alshawi
AT&T Labs
180 Park Avenue
Florham Park, NJ 07932-0971, USA
hiyan@research.att.com

Fei Xia
Department of Computer and
Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
fxia@cis.upenn.edu

Abstract

We describe the head transducer model used in an experimental English-to-Mandarin speech translation system. Head transduction is a translation method in which weighted finite state transducers are associated with source-target word pairs. The method is suitable for speech translation because it allows efficient bottom up processing. The head transducers in the experimental system have a wider range of output positions than input positions. This asymmetry is motivated by a tradeoff between model complexity and search efficiency.

1 Introduction

In this paper we describe the *head transducer* model used for translation in an experimental English-to-Mandarin speech translation system. Head transducer models consist of collections of weighted finite state transducers associated with pairs of lexical items in a bilingual lexicon. Head transducers operate "outwards" from the heads of phrases; they convert the left and right dependents of a source word into the left and right dependents of a corresponding target word.

The transducer model can be characterized as a statistical translation model, but unlike the models proposed by Brown et al. (1990, 1993), these models have non-uniform linguistically motivated structure, at present coded by hand. The underlying linguistic structure of these models is similar to dependency grammar (Hudson 1984), although dependency representations are not explicitly constructed in our approach to translation. The original motivation for the head transducer models was that they are simpler and more amenable to automatic model structure acquisition as compared with earlier transfer models.

We first describe the head transduction approach in general in Section 2. In Section 3 we explain properties of the particular head transducers used in the experimental English-to-Mandarin speech translator. In Section 4, we explain how head transducers help satisfy the requirements of the speech translation application, and we conclude in Section 5.

2 Bilingual Head Transduction

2.1 Bilingual Head Transducers

A head transducer M is a finite state machine associated with a pair of words, a source word w and a target word v. In fact, w is taken from the set V_1 consisting of the source language vocabulary augmented by the "empty word" ϵ, and v is taken from V_2, the target language vocabulary augmented with ϵ. A head transducer reads from a pair of source sequences, a left source sequence L_1 and a right source sequence R_1; it writes to a pair of target sequences, a left target sequence L_2 and a right target sequence R_2 (Figure 1).

Head transducers were introduced in Alshawi 1996b, where the symbols in the source and target sequences are source and target words respectively. In the model described in this paper, the symbols written are dependency relation symbols, or the empty symbol ϵ. The use of relation symbols here is a result of the historical development of the system from an earlier transfer model. A conceptually simpler translator can be built using head transducer models with only lexical items, in which case the distinction between different dependents is implicit in the state of a transducer. In head transducer models, the use of relations corresponds to a type of class-based model (cf Je-

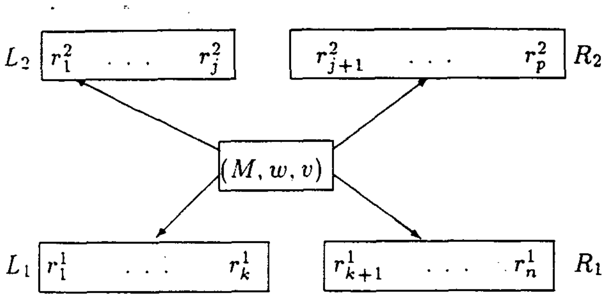
L_2
$r_1^2 \; . \; . \; . \; r_j^2$
$r_{j+1}^2 \; . \; . \; . \; r_p^2$
R_2
(M, w, v)
L_1
$r_1^1 \; . \; . \; . \; r_k^1$
$r_{k+1}^1 \; . \; . \; . \; r_n^1$
R_1

Figure 1: Head transducer M converts the sequences of left and right relations $\langle r_1^1 \ldots r_k^1 \rangle$ and $\langle r_{k+1}^1 \ldots r_n^1 \rangle$ of w into left and right relations $\langle r_1^2 \ldots r_j^2 \rangle$ and $\langle r_{j+1}^2 \ldots r_p^2 \rangle$ of v.

linek, Mercer and Roukos, 1992).

We can think of the transducer as simultaneously deriving the source and target sequences through a series of transitions followed by a stop action. From a state q_i these actions are as follows:

- Left transition: write a symbol r_1 onto the right end of L_1, write symbol r_2 to position α in the target sequences, and enter state q_{i+1}.
- Right transition: write a symbol r_1 onto the left end of R_1, write a symbol r_2 to position α in the target sequences, and enter state q_{i+1}.
- Stop: stop in state q_i, at which point the sequences L_1, R_1, L_2 and R_2 are considered complete.

In simple head transducers, the target positions α can be restricted in a similar way to the source positions, i.e., the right end of L_2 or the left end of R_2. The version we used for English-to-Chinese translation allows additional target positions, as explained in Section 3.

2.2 Recursive Head Transduction

We can apply a set of head transducers recursively to derive a pair of source-target ordered dependency trees. This is a recursive process in which the dependency relations for corresponding nodes in the two trees are derived by a head transducer. In addition to the actions performed by the head transducers, this derivation process involves the actions:

- Selection of a pair of words $w_0 \in V_1$ and $v_0 \in V_2$, and a head transducer M_0 to start the entire derivation.
- Selection of a pair of dependent words w' and v' and transducer M' given head words w and v and source and target dependency relations r_1 and r_2. ($w, w' \in V_1$; $v, v' \in V_2$.)

The recursion takes place by running a head transducer (M' in the second action above) to derive local dependency trees for corresponding pairs of dependent words $\langle w', v' \rangle$. In practice, we restrict the selection of such pairs to those provided by a bilingual lexicon for the two languages. This process of recursive transduction of local trees is shown graphically in Figure 2 in which the pair of words starting the entire derivation is $\langle w4, v4 \rangle$.

2.3 Translator

A translator based on head transducers consists of the following components:

- A bilingual lexicon in which entries are 5-tuples $\langle w, v, M, q, c \rangle$, associating a pair of source-target words with a head transducer M, an initial state q, and a cost c.
- A parameter table giving the costs of actions for head transducers and the recursive transduction process.
- A transduction search engine for finding the minimum cost target string for an input source string (or recognizer speech lattice). The search algorithm used in our implementation is a head-outwards dynamic programming algorithm similar to the parsing algorithm for monolingual head acceptors described in Alshawi 1996a. Head-outwards processing techniques were developed orgninally for lexically-driven parsing (Sata and Stock 1989, Kay 1989).

3 English-Chinese Head Transducers

3.1 Source and Target Positions

In deciding the set of allowable positions for source and target transitions, there are tradeoffs involving model size, flexibility for modeling word-order changes in translation, and computational efficiency of the search for lowest cost transductions.

These tradeoffs led us to constrain the source positions of transitions to just two, specifically the simple left and right source positions mentioned in the description of transitions in Section 2.1. This restriction means that the transduction search can be carried out with the type of algorithm used for

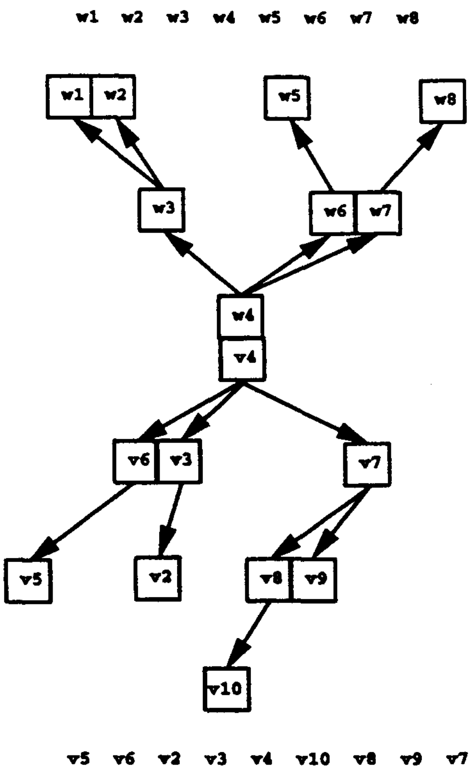

Figure 2: Recursive head transduction of a string

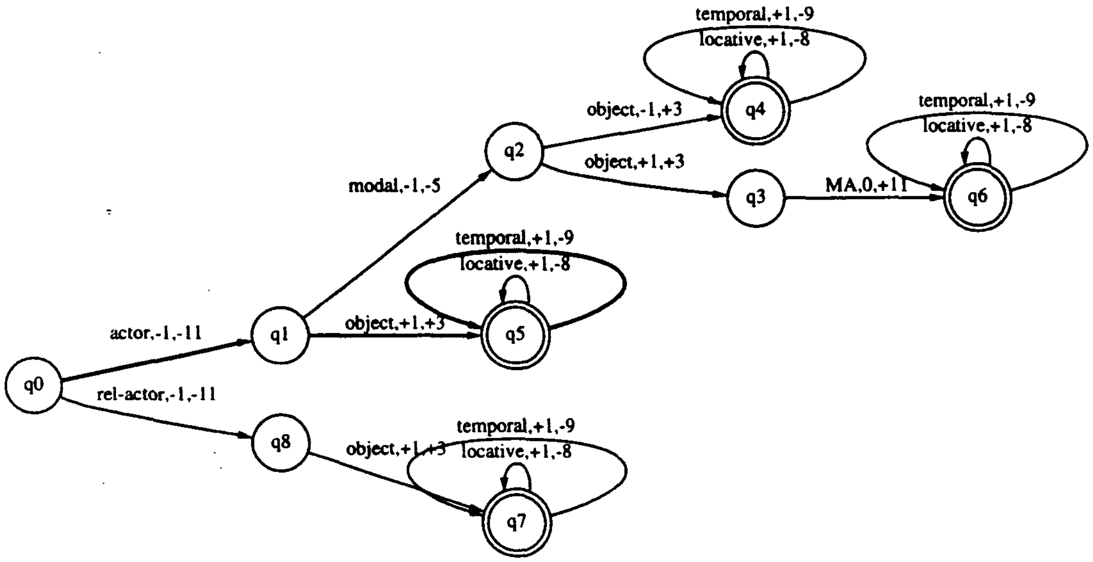

Figure 3: Simplified transitive verb head transducer

head-outwards context free parsing. In particular, we use a dynamic programming tabular algorithm to find the minimal cost transduction of a word string or word-lattice from a speech recognizer. The algorithm maintains optimal "active-edges" spanning a segment of the input string (or two states in the recognition word-lattice). This use of context free algorithms is not possible if the number of possible source positions for transductions is increased so that incomplete transducer source sequences are no longer simple segments.

However, the number of target positions for transductions is not constrained by these efficiency considerations. For English-to-Chinese translation, we can decrease the complexity of the transducers (i.e. reduce the number of states and transitions they have) by allowing multiple target positions to the left and right of the head. The motivation for this is that the required reordering of dependents can be achieved with fewer transducer states by accumulating the dependents into subsequences to the left and right of the head. The actual left and right target sequences are formed by concatenating these subsequences. We can use the following notation to number these additional positions. The head is notionally at position 0, and the "standard" positions immediately to the left and right of the head are numbered as -1 and +1 respectively. The position that extends the kth subsequence to the left of the head outwards from the head is numbered $-2k+1$, while the position that extends this same subsequence inwards towards the head is labeled $-2k$. The positions to the right of the head are numbered analogously with positive integers.

3.2 Examples of Dependency Relation Head Transducers

An example of the structure of a simplified head transducer for converting the dependents of a typical English transitive verb into those for a corresponding Chinese verb is shown in Figure 3. The nodes in the figure correspond to states; a bilingual lexical entry would specify $q0$ as the initial state in this case. Transitions are shown as arcs between states; the label on an arc specifies the relation symbol, source position, and target position, respectively. Stop actions are not shown, though states allowing stop actions are shown as double circles, the usual convention for final states. A typical path through the state diagram is shown in bold: this converts the English dependency sequence for statement sentences with the pattern

```
actor head object temporal
```

into the corresponding Chinese sequence

```
actor temporal head object.
```

Similarly, an English dependency sequence for yes-no questions

```
modal actor head object temporal
```

is converted into the Chinese sequence

```
actor temporal modal head object MA,
```

the transducer stopping in state $q6$, MA being the relation between the head verb and the Chinese particle for yes-no questions. The final states for this transducer network are kept distinct so that different costs can be assigned by training to the stop actions and modifier transitions at these states.

Another example is the English-to-Chinese head transducer for noun phrase dependency relations shown in Figure 4. Typical target positions for transitions corresponding to noun phrase modification (noun phrases are head-final in Chinese) are as follows:

head:	0	(flight)
nominal:	-1	(airline)
adjective:	-3	(cheap)
possessive:	-5	(Continental's)
relative:	-6	(that leaves NYC)
locative:	-8	(from NYC)
temporal:	-9	(before one pm)
classifier:	-10	(pint)
specifier:	-11	(all)
cardinal:	-11	(five)
ordinal:	-11	(first)
DE:	-2, -4, or -6	

The position for transitions emitting the Chinese particle pronounced DE may be either -2, -4, or -6, depending on the transducer states for the transition. The different states effectively code the presence of different modifier types. It should also be noted that the above positions do not completely define the order of modifiers in the transduction. For example, the relative order of target specifiers, cardinals, and ordinals will depend on the order of these modifiers in the source.

3.3 Model Construction

The head transducer model was trained and evaluated on English-to-Mandarin Chinese translation of transcribed utterances from the ATIS corpus (Hirschman et al. 1993). By training here we

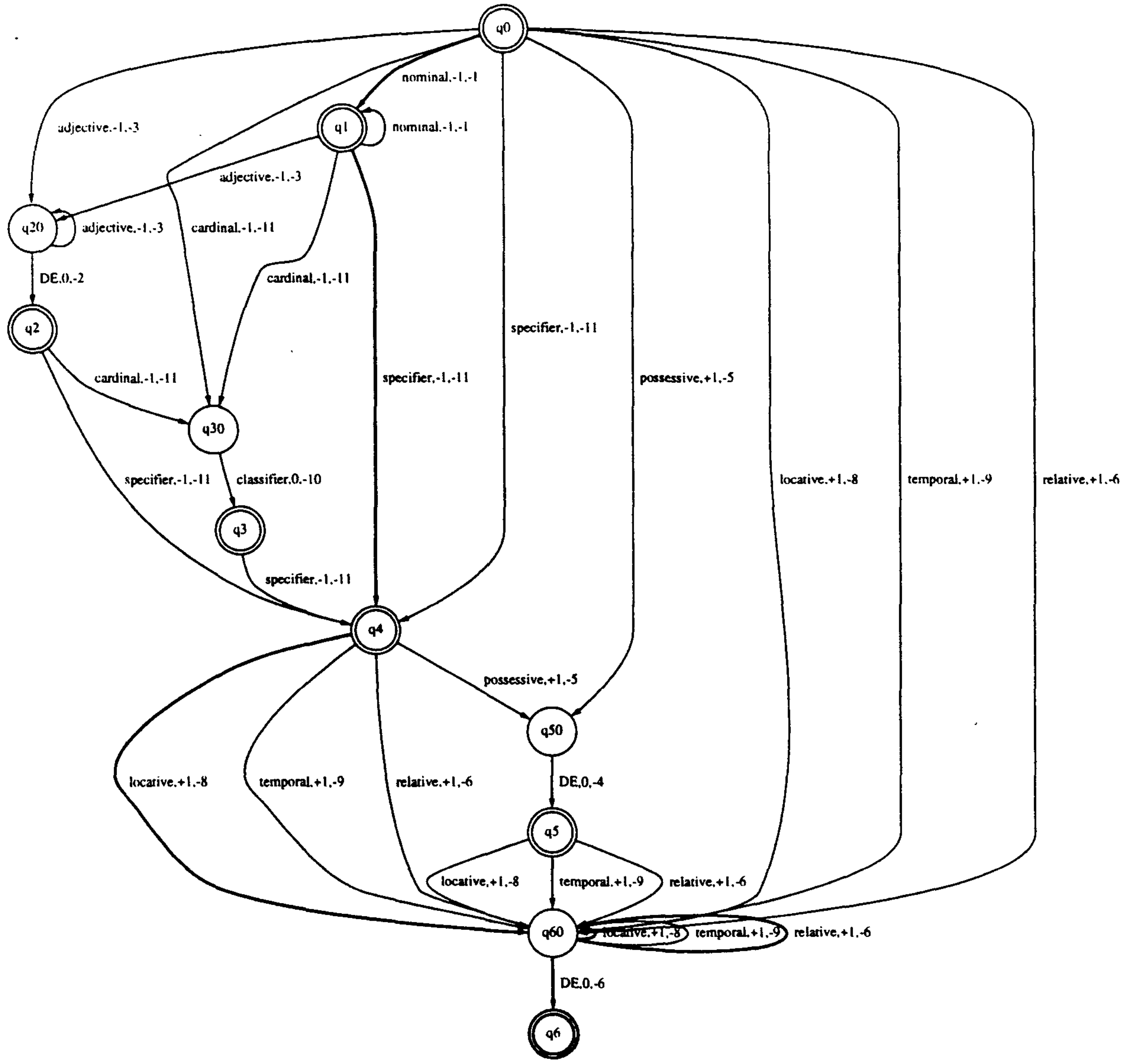

Figure 4: Head transducer for noun phrase dependents

simply mean assignment of the cost functions for fixed model structures. These model structures were coded by hand as a head transducer lexicon.

The head transducers were built by modifying the English head acceptors defined for an earlier transfer-based system (Alshawi 1996a). This involved the addition of target relations, including some epsilon relations, to automaton transitions. In some cases, the automata needed to be modified to include additional states, and also some transitions with epsilon relations on the English (source) side. Typically, such cases arise when an additional particle needs to be generated on the target side, for example the yes-no question particle in Chinese. The inclusion of such particles often depended on additional distinctions not present in the original English automata, hence the requirement for additional states in the bilingual transducer versions.

In fact, many of the automata in these entries had the same structure, and are independent of the ATIS domain. Domain dependence and the differences in word behavior (for example the differences in behavior between two verbs with the same subcategorization) were due to the costs applied when running the automata. The method used to assign the cost parameters for the model can be characterized as "supervised discriminative training". In this method, costs are computed by tracing the events involved in producing translations of sentences from a source training corpus; a bilingual speaker classifies the output translations as positive or negative examples of acceptable translations. Details of this cost assignment method are presented in Alshawi and Buchsbaum 1997.

4 Head Transducers in Speech Translation

Speech translation has special requirements for efficiency and robustness. We believe that head transduction models have certain advantages that help satisfy these requirements.

Ranking Head transduction models are weighted, so the costs for translation derivations can be combined with those from acoustic processing. Weighted models can also contribute to efficiency because dynamic programming can be used to eliminate suboptimal derivations. This is particularly important when the input is in the form of word lattices. Since the contributions of both the source, target, and bilingual components of the models are applied simultaneously when computing the costs of partial derivations, there is no need to pass multiple alternatives forwards from source analysis to transfer to generation; the translation ranked globally optimal is computed with a single admissible search.

Efficiency In addition to the points made in the preceding paragraph on ranking, we noted earlier that transduction with appropriately restricted source positions for transitions can be carried out with search techniques similar to context free parsing (e.g. Younger 1967). Head outward processing with a lexicalized model also the obvious advantage to efficiency that only the part of the model related to the source words in the input needs to be active during the search process. In an experiment comparing the efficiency of head transduction to our earlier transfer approach, the average time for translating transcribed utterances from the ATIS corpus was 1.09 seconds for transfer and 0.17 for head transduction. This speed improvement was possible while also improving memory usage and translation accuracy. Details of the experiment are presented in Alshawi, Buchsbaum, and Xia, 1997. The efficiency of head transduction has allowed us to start experimenting with (pruned) word lattices from speech recognition with the aim of producing translations from such word lattices in real time.

Robustness Bottom-up lexicalized translation is inherently more robust than top-down processing since it allows maximal incomplete partial derivations to be identified when complete derivations are not possible. This is particularly important in the case of speech translation because the input string or word lattice often represents fragmentary, illformed, or "after thought" phrases. When complete derivations are not possible, our experimental system searches for a span of the input string or lattice with the fewest fragments (or the lowest cost such span if there are several). Lowest-cost translations of such fragments will already have been produced by the transduction algorithm, so an approximate translation of the utterance can be formed by concatenating the fragments in temporal order. In the limit, this approach degrades gracefully into word-for-word translation with the most likely translation of each input word being selected.

5 Conclusion

Head transducers offer efficiency and robustness advantages to the speech translation application; there is empirical evidence supporting this claim at least in the case of comparison with a transfer approach. We have also argued that allowing multiple target positions for transitions increases the flexibility of transducers without an adverse effect on efficiency. The focus of our current research is to take advantage of the relative simplicity of head transducer models in working towards fully automatic model acquisition.

References

Alshawi, H., A.L. Buchsbaum, and F. Xia. 1997. "A Comparison of Head Transducers and Transfer for a Limited Domain Translation Application". In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid.

Alshawi, H. and A.L. Buchsbaum. 1997. "State-Transition Cost Functions and an Application to Language Translation". In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Munich, Germany.

Alshawi, H. 1996a. "Head Automata and Bilingual Tiling: Translation with Minimal Representations". In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, 167-176.

Alshawi, H. 1996b. "Head Automata for Speech Translation". In *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, Pennsylvania.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer and P. Rossin. 1990. "A Statistical Approach to Machine Translation". *Computational Linguistics* 16:79–85.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics* 19:263–312.

Chen, K.H. and H. H. Chen. 1992. "Attachment and Transfer of Prepositional Phrases with Constraint Propagation". *Computer Processing of Chinese and Oriental Languages*, Vol. 6, No. 2, 123–142.

Hudson, R.A. 1984. *Word Grammar*. Blackwell, Oxford.

Hirschman, L., M. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallett, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann. 1993. "Multi-Site Data Collection and Evaluation in Spoken Language Understanding". In *Proceedings of the Human Language Technology Workshop*, Morgan Kaufmann, San Francisco, 19–24.

Jelinek, F., R.L. Mercer and S. Roukos. 1992. "Principles of Lexical Language Modeling for Speech Recognition". In S. Furui and M.M. Sondhi (eds.), *Advances in Speech Signal Processing*, Marcel Dekker, New York.

Kay, M. 1989. "Head Driven Parsing". In *Proceedings of the Workshop on Parsing Technologies*, Pittsburgh, 1989.

Sata, G. and O. Stock. 1989. "Head-Driven Bidirectional Parsing". In *Proceedings of the Workshop on Parsing Technologies*, Pittsburgh.

Younger, D. 1967. Recognition and Parsing of Context-Free Languages in Time n^3. *Information and Control*, 10, 189–208.