NRC·CNRC

Institute for Information Technology

MT: The Current Research Landscape

Roland Kuhn and Pierre Isabelle

August 2009

Goals & Trends in Presentation

- Two goals for this talk:
  1. Provide overview of current research in MT.
  2. Provide overview of research papers at this conference.
- Trends & background information:
  - More & more research activity
  - most current research in MT involves statistical MT = SMT (as opposed to rule-based MT = RBMT)
    - open-source packages & data have lowered barriers to entry
    - *e.g.*, GIZA++ for word alignment, Moses decoding and LDC for data
  - SMT needs bilingual training data - much research on gathering such data
  - tuning SMT system requires automatic evaluation metrics – you'll hear "BLEU" a lot
  - MT teams participate in regular international competitions (e.g., NIST, ACL)

Goals & Trends in Presentation

- Trends & background information (cont.):
  - Funding of research:
    - US gov't interested mainly in En as target (GALE = Ar → En, Ch → En; NIST = same + Urdu → En);
    - EU mostly interested in European languages;
    - Large American corporations (e.g. Microsoft) interested mainly in En as source.
  - SMT systems are *quickly improving* (better algorithms, more training data)
  - Some European language pairs (En ↔ Fr, En ↔ Sp) may have reached quality required for wide usability
  - Ch↔ En increasingly important; more & more Chinese researchers getting involved
  - More & more use of syntax in SMT
  - Combination of MT systems is surprisingly effective
  - Google Translate's SMT has become the gold standard; being used surreptitiously by professional translators, kids cheating on homework, *etc.*
  - Commercial offerings available for deploying SMT in-house

Some Gaps

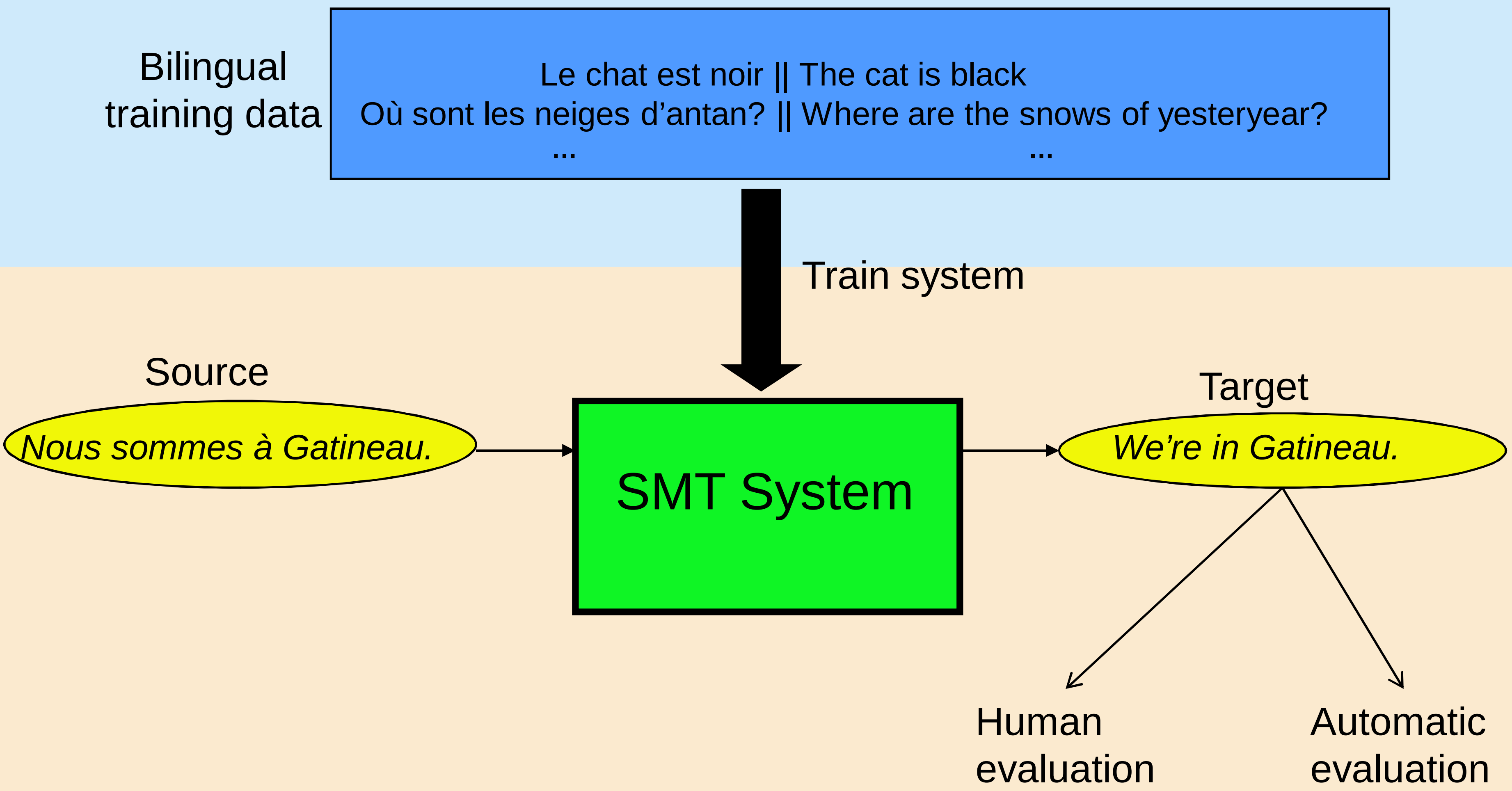
- Not enough user studies
- Not enough work on incorporating MT into translators' tools (*e.g.*, translation memory)
- Too much focus on clever new techniques applied to old problems, instead of known techniques applied to new problems? (Richard Sproat)
- Not enough work on morphologically rich languages
- Too much focus on language pairs where either the source or target language is English?
  - *E.g.*, EACL 2009 workshop evaluated Fr ↔ En, Sp ↔ En, De ↔ En, Cz ↔ En, Hu ↔ En)

Themes of Presentation & Research Programme

1. MT-based tools
2. Evaluation of MT systems
3. Multilingual issues
4. Training corpora & data mining
5. SMT system training & decoding*****
6. System combination, system adaptation, & new types of MT*****
7. Syntax & reordering in SMT systems*****

***** = require some knowledge of internal workings of SMT

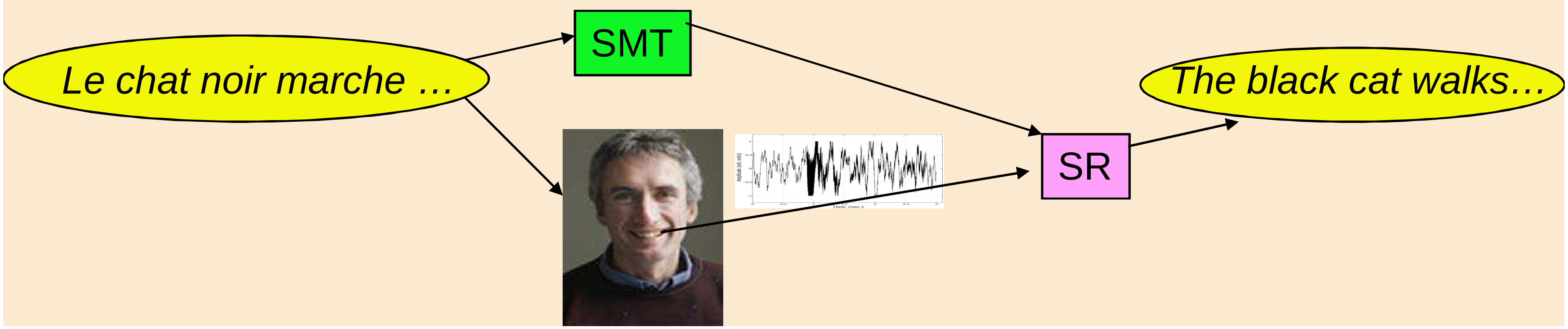
What you need to know about SMT (for first part of presentation)



1. MT-based tools

Making human translators more productive

- (Koehn & Haddow): three options for translators: 1) suggestions for sentence completion; 2) word & phrase translation options, and 3) post-editing of MT. Fr → En user study looks at productivity impact and users' impressions.
- (Simard & Isabelle): study of several different ways of integrating SMT into a translation memory, to create a hybrid that's better than either. A component filters out the low-quality SMT output (confidence estimation).
- (Specia et al.): confidence estimation for SMT using machine learning.
- (Reddy et al.): when translator is dictating translation to speech recognition (SR) system, use SMT to help SR. With help of named entity recognizer, attained up to 32% decrease in word error rate.



1. MT-based tools

Making human translators more productive (cont'd)

- (Huet *et al*.): To improve bilingual concordancer, find all translations of source phrase in bitext, using a 2-pass algorithm based on IBM2 (from word alignment step in training of SMT systems).

MT for dialogue

- (Starlander & Estrella): MedSLT is speech dialogue system for multilingual doctor-patient communication, with back translation. Grammar-based MT using Interlingua, with help module that guides users towards covered domain. Evaluation of SR performance, MT performance, & usability.
- (Zhang): Proposes SMT for translation of chat messages in Second Life; plan to build context-awareness into SMT model.

Improving text written in 2nd language

- (Désilets & Hermet) L2 sentences (French written by anglophones) translated by Google Fr → En and then backtranslated by Google En → Fr. Surprisingly, errors in L2 often get repaired (see evaluation part of paper).

NRC-CNRC Institute for Information Technology

2. Evaluation of MT systems Background

- Various kinds of human evaluations depending on how MT is being used:
    - Subjective adequacy/fluency assessments
    - Productivity measurements
    - Comprehension tests based on MT output, *etc*.
- Automatic evaluation of MT involves measuring similarity of MT output to 1 or more reference translations. Obvious flaw:
    - Ref = « The man spoke rudely to me »
    - Output 1 = « The man spoke politely to me »
    - Output 2 = « He was insolent »
    - → output 1 will score higher.
- Automatic metrics used in developing SMT systems
    - Compare thousands of variants of each system → far too much work for humans!
- Commonly used automatic metrics:
    - BLEU (comparison of n-gram matches between MT output and ref.), NIST (similar to BLEU),
    - METEOR (takes into account stemming & synonymy),
    - TER (related to edit distance), *etc.*
- Problem: for some reason, automatic metrics seem to favour SMT over RBMT

NRC-CNRC
Institute for Information Technology

2. Evaluation of MT systems

New human evaluation methods

- (Ogden *et al*.): evaluate quality of Cross-Language Instant Messaging by having one user question another about photo being shown on screen; the faster the correct photo selected, the better the MT quality.
- (Doherty & O'Brien): native speakers of target language read MT output, and their eyes are tracked. Gaze time shorter for high-quality sentences. Maybe eye tracking is faster & more objective than subjective adequacy/fluency?

Automatic evaluation methods

- (Tatsumi): looks at correlation between several automatic metrics and postediting speed for English → Japanese;
- (Zhao *et al*.): look at results of CWMT2008 evaluation, focusing on 2 new metrics: BLEU-SBP and linguistic check-point method;
- (Condon *et al*.): shows how automatic metrics overestimate difficulty of MT into Arabic, & how they can be fixed.

3. Multilingual issues

What makes a language pair hard for SMT?

- According to (Birch *et al*., EMNLP 2008), 3 strong predictors:
  1. amount of reordering
  2. morphological complexity of T
  3. relatedness of S & T.

  Each one accounts for 1/3 of variation in BLEU (3/4 together), for 110 European language pairs.
- (Koehn et al.): extend this work to 462 European language pairs.
  - Also add another explanatory factor, *entropy*.
  - Paper also looks at translating via a pivot language & multisource SMT.
- (Rayner et al.): use artificial data provided by a RBMT system to assess quality of translations for different language pairs
  - E.g. En ↔ Fr much easier than {En, Fr} → Ja.
- Still lots of work needed here, esp. on non-European languages
  - E.g. why is Ar → En so much easier than Ch → En?
  - Dekai Wu's hypothesis: lots of Europe → Middle East cultural links (panel talk, DARPA GALE meeting, Apr. 2008)

3. Multilingual issues

How can we handle a low-resource language pair?

- (Genzel *et al*.): work on En Yiddish.
  Yiddish is a Germanic language with borrowings from Polish & Hebrew, written in the Hebrew alphabet.
  Authors cleverly use bridging information from German, Polish, & Hebrew to learn meanings of cognates.
- (Varga & Yokoyama): for Japanese → Hungarian, build RBMT system automatically by learning syntactic transfer rules from a parsed bilingual corpus and a bilingual dictionary.

4. Training corpora & data mining Background

- To train SMT systems or build multilingual terminology databases, we need sentence-aligned bilingual text
- Traditionally, SMT researchers have used data produced or collected by governments, or by LDC: Canadian Hansard, Hong Kong Hansard, Europarl, UN corpus, *etc*.
- (Resnik and Smith, « The Web as a parallel corpus », Computational Linguistics, 2003): proposed programs that mine the Web for parallel text.
- In using the Web as data source, one often encounters *comparable* corpora: pairs of texts that are not exact translations of each other, but that cover the same semantic material.

  → These too can be useful in training SMT systems.

4. Training corpora & data mining Papers at MT Summit

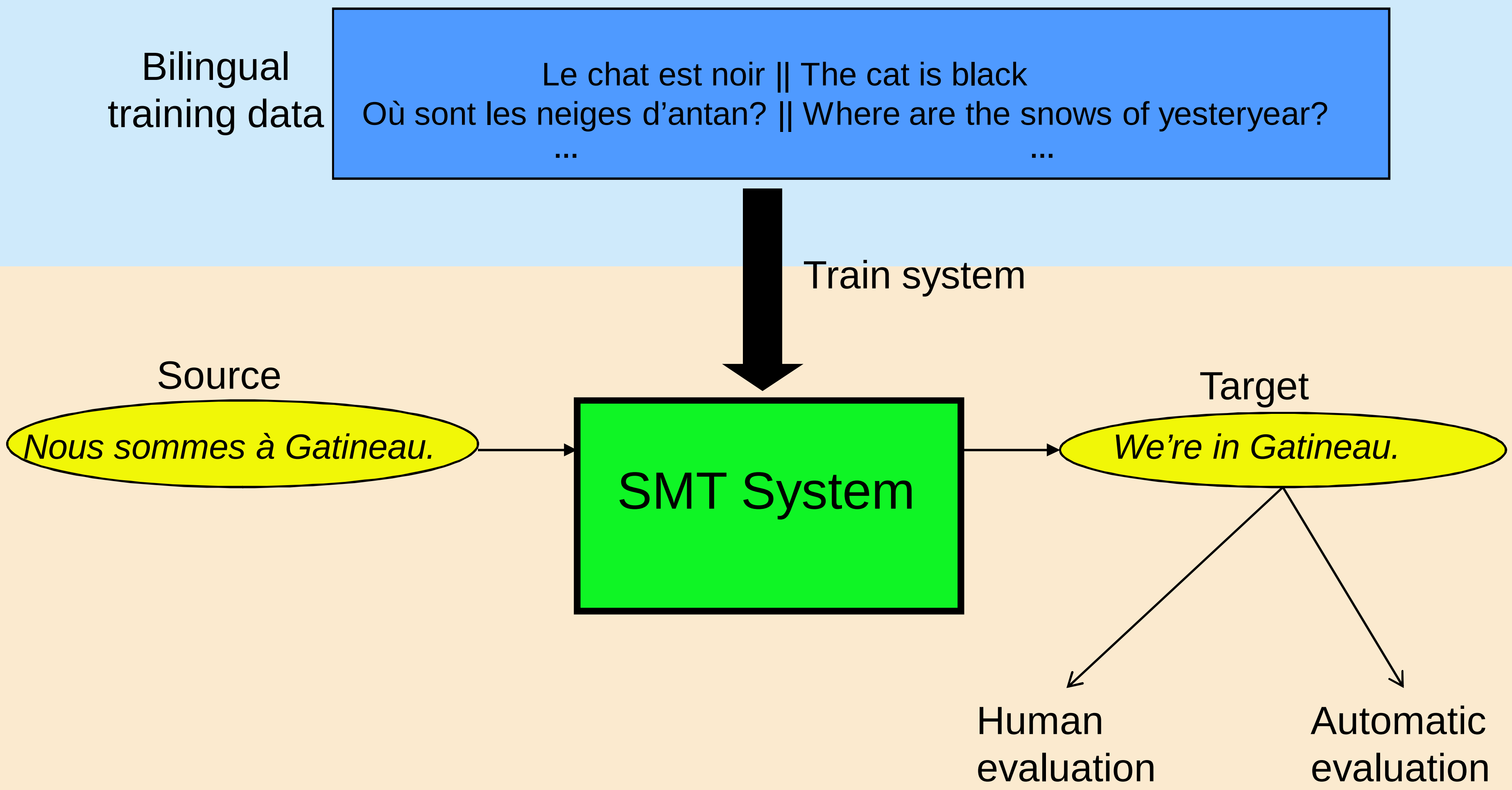
- (Rafalovitch & Dale): describes a parallel corpus gathered from official resolutions of the UN, in paragraph-aligned En, Fr, Sp, Ru, Ch & Ar. About 3 million words per language.
- (Yu & Tsujii): use Wikipedia as a source of comparable corpora, then extract bilingual dictionary from the comparable corpora.
- (Prochasson *et al*.): extract bilingual lexica (En ↔ Ja, Fr ↔ Ja) from comparable corpora.
- (Ishisaka *et al*.): create an En-Ja parallel corpus from open source software manuals on the Web.
- (Utiyama, Kawahara *et al*.): extract parallel sentences from mixed-language Web pages.
- (Zhu *et al*.): detailed description of extracting aligned sentences from Web data.

Two Outliers

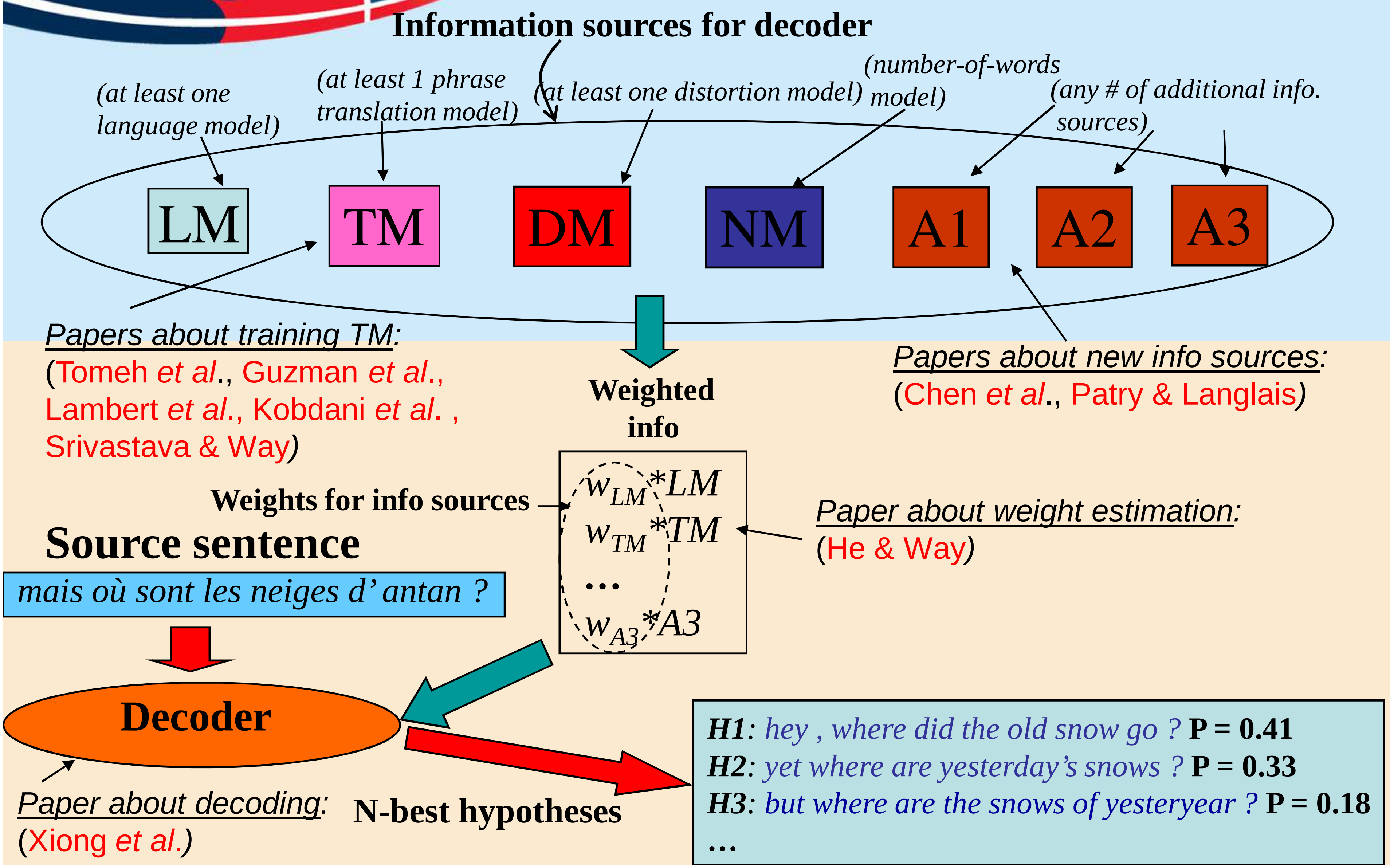
- (Kurokawa et al.): shows that it's possible to detect which half of bitext is original, which translated (90% document accuracy); also show it's better to train SMT system on bilingual data that has same direction as desired task.
- (Utiyama, Abekawa et al.): describes a site that hosts online volunteer translators.

NRC·CNRC
Institute for Information Technology

What you needed to know about SMT (for first part of presentation)



What you need to know about phrase-based SMT (for 2nd part of presentation)



5. SMT system training and decoding
The phrase translation model

Phrase-based approach introduced around 1998 by Franz Josef Och & others (Ney, Wong, Marcu)

Example: **« cul de sac »**

word-based translation = **« ass of bag »** (N. Am), **« arse of bag »** (British)
phrase-based translation = **« dead end »** (N. Am.), **« blind alley »** (British)

This knowledge is stored in a phrase table: collection of conditional probabilities of form P(S|T) = *backward* phrase table or P(T|S) = *forward* phrase table.

backward: P(S|T)

p(sac|bag) = 0.9
p(sacoche|bag) = 0.1
…
p(cul de sac|dead end) = 0.7
p(impasse|dead end) = 0.3
…

forward: P(T|S)

p(bag|sac) = 0.5
p(hand bag|sac) = 0.2
…
p(ass|cul) = 0.5
p(dead end|cul de sac) = 0.85
…

NRC·CNRC
Institute for Information Technology

5. SMT system training and decoding
Training the phrase translation model

Paper about word alignment:
(Kobdani *et al.*)

Bilingual sentence-aligned corpus

I want to go home. I saw him on television. …	Je veux aller chez moi. Je l'ai vu à la télévision. …

Word alignment (via IBM or HMM models)

Je l'ai vu à la télévision.
I saw him on television.

Phrase extraction

Papers about phrase extraction:
(Guzman *et al.*, Lambert *et al.*, Srivastava & Way)

***(Je, I)**, **(Je l' ai vu, I saw him)**,*
***(ai vu, saw)**, **(l' ai vu à la, saw him on)**,*
…

Phrase table creation

P(S|T)
p(je|I) = 0.93
p(ai vu|saw) = 0.6
p(sac|bag) = 0.8
p(sacoche|bag) = 0.1
…

(optional) Phrase table pruning

Paper about phrase table pruning:
(Tomeh *et al.*)

NRC-CNRC
Institute for Information Technology

5. SMT system training and decoding
Training the phrase translation model

Papers about training TM:

- (Kobdani *et al*.): new *m-to-n* word alignment heuristic, which works better than IBM1 in terms of F-measure (background: Och & Ney, "A Comparison of Alignment Models for Statistical Machine Translation", COLING 2000).
- (Srivastava & Way): try 3 different syntactic methods for extracting phrases – none as good on its own, but all helpful when used as a complement to standard (non-syntactic) approach (experiments on *Fr→En* Europarl)
- (Guzman *et al*., Lambert *et al*.): analyze relationship between word alignment and phrase extraction: fewer word links → more phrase pairs. (Guzman *et al*.) shows more word links → higher quality phrase pairs. Using # of unaligned words in phrase pairs as info source for decoding → +2 BLEU (on large Ch → En task).
- (Tomeh *et al*.): drastic pruning of phrase table through significance testing. Statistical criterion: « noise » instead « p-value ». Large decrease in table size AND greater BLEU gains.

  (Background: Johnson *et al*., « Improving Translation Quality by Discarding Most of the Phrasetable », Proc. EMNLP-CoNLL, 2007)

5. SMT system training and decoding
New info sources & weight estimation

Papers about new info sources:

- (Chen *et al*.): Use four measures of association between phrases **s** and **t**, reflecting how often a sentence with **s** was aligned with a sentence with **t**, → +0.5 – 0.6 BLEU individually, +0.6 – 0.7 BLEU together (large Ch→En task).
- (Patry & Langlais): use a multilayer perceptron to predict target words from source words (using only sentence alignments, not word alignments).

Paper about weight estimation:

- Background: weights on info sources have huge impact on performance. Standard MERT technique for estimating weights is (F. Och, « Minimum error rate training in statistical machine translation », ACL 2003).
- (He & Way) argue that MERT works better if you use a mix of metrics, rather than just one (*e.g.*, BLEU).

5. SMT system training and decoding
Decoding process

<u>Order</u>: Target hypotheses grow left->right, from source segments consumed in any order

Source: **s1 s2 s3 s4 s5 s6 s7 s8 s9**

Segmentation

Backward Table

<u>P(S|T)</u>

p(s2 s3 | t8)

p(s2 s3 | t5 t3)

…

p(s3 s4 | t4 t9)

…

<u>phrase table</u>:

1. suggests possible segments
2. supplies phrase translation scores

(pick s2 s3 first)

Source: **s1 [s2 s3] s4 s5 s6 s7 s8 s9**

(phrase transl)

Tgt hyp: **t8| …**

Tgt hyp: **t5 t3| …**

…

(pick s5 s6 s7)

Source: **s1 [s2 s3] s4 [s5 s6 s7] s8 s9**

(phrase transl)

Tgt hyp: **t8| t6 t2| …**

…

(pick s3 s4 first)

Source: **s1 s2 [s3 s4] s5 s6 s7 s8 s9**

(phrase transl)

Tgt hyp: **t4 t9| …**

…

NRC-CNRC
Institute for Information Technology

5. SMT system training and decoding
Paper about decoding process

Paper about more efficient decoding:

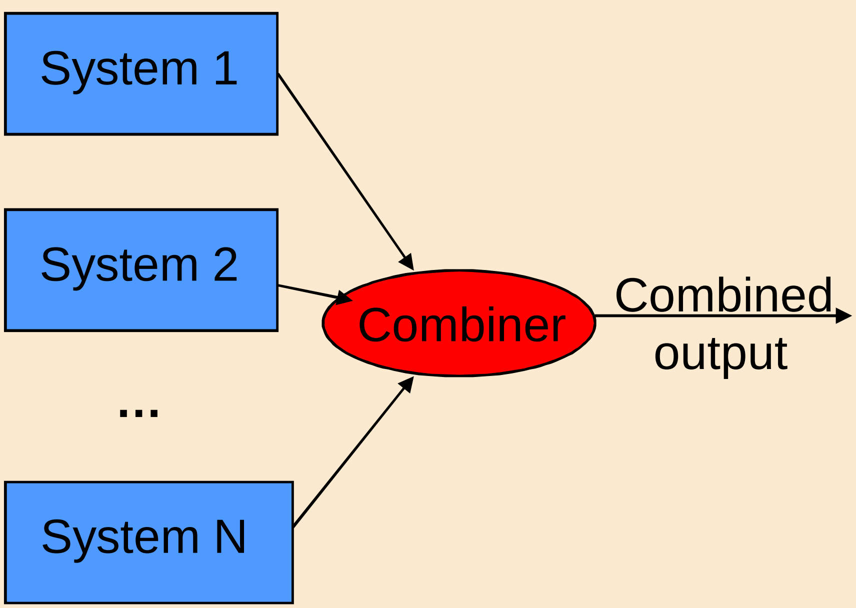
- Much recent research into ways of speeding up decoding
  - *e.g.*, work on cube pruning (Huang & Chiang, « Forest Rescoring: Faster Decoding with Integrated Language Models », ACL 2007).
- One widely used method is beam thresholding, where only hypotheses with score >   * (score of best hypothesis) are retained.
- (Xiong *et al*.) propose two variations on beam thresholding that lead to major speedup with no decline in BLEU; the first variation yields a speedup even when cube pruning applied.

6. System combination, system adaptation, & new types of MT

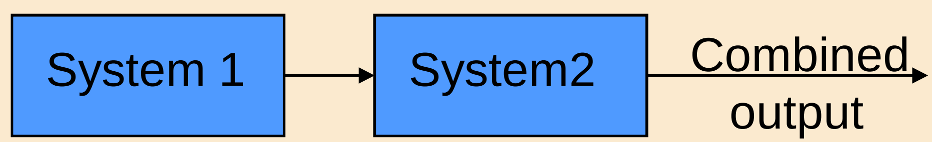
Background:

- Recently, excellent results from system combination. *E.g.*, NIST08 constrained track Ch →En: highest score from parallel **combination of eight systems** = 30.9 BLEU; best of eight systems has 26.2 BLEU
  (He *et al*., « Indirect-HMM-based Hypothesis Alignment … », EMNLP 2008).
- Two common kinds of system combination:

Parallel system combination



Serial system combination



NRC·CNRC
Institute for Information Technology

6. System combination, system adaptation, & new types of MT (cont.)

Papers:

- (Thurmair) Overview of 3 kinds of hybrid MT systems: 1. coupled (parallel or serial combination); 2. predominantly RBMT or SMT with peripheral elements of the other approach; 3. genuinely hybrid. Also discusses domain adaptation.
- (Du & Way): Typical parallel combination: a) align MT outputs together; b) build confusion network; and c) select consensus hypothesis. This paper: align source with each MT output to help build confusion network. On En → Fr task, +0.2 BLEU over baseline parallel combination; on Ch → En task, +0.6 BLEU.
- (Aikawa & Ruopp): Serially combine *syntax(treelet)-based* system with *phrase-based* system. For three language pairs (En →Sp, En → De, En →Ja) proves better than either constituent system by 1.0 – 3.7 BLEU. Paper contains analysis of improvements: more fluency, better handling of inflections.
  Background in (Simard *et al*., « Statistical Phrase-based Post-editing », NAACL-HLT 2007).

6. System combination, system adaptation, & new types of MT (cont.)

Two papers on adaptation:

- (Schwenk & Senellart) adapt a generic Ar → Fr system (trained on UN data) to the news domain by self-training on Arabic news data (from Arabic Gigaword); +3.5 BLEU.
- (Dugast *et al*.) adapt an En → Fr RBMT system by adding to its phrasal lexicon 67K phrase pairs extracted from a bilingual corpus (Europarl) via SMT-like methods; +3 BLEU.

Two new approaches:

- (Kamatani *et al*.) Ja → En system; syntactic rules split source sentence into segments; each segment translated by appropriate method (EBMT or RBMT).
- (Soderland *et al*.) «Lemmatic MT »: focus only on adequacy for low-resource language pairs, forget fluency & syntax. Resources: only need a bilingual dictionary. Big problem: polysemy; handle through back-translation & word sense disambiguation.
  Claim broader language coverage than Google MT and often better adequacy for languages Google does cover.

7. Syntax & reordering in SMT systems (A) Word ordering in PBSMT

- Phrase pairs recorded in training capture much local ordering
  - <*the small cat*, *le petit chat*>
  - <*the black cat*, *le chat noir*>
- Phrase reordering through distance-based « distortion »
  - Let phrases move around individually; try many different orderings
  - *Distance-based score*: bonus for keeping close/far words that were close/far in SL
  - Target-language *LM score*: big bonus for word order that increases a priori probability of TL (fluidity bias!)
- In more recent PBSMT models, lexical conditioning is added:
  - Phrase pairs assigned an ordering type wrt previously translated element
    - Monotone: keep going in same direction as previous element
    - Swap: swap order with previous element
    - Discontinuous: send away from previous element

7. Syntax & reordering in SMT systems (B) Ordering problems in PBSMT

- Lack of generalization

OK: *The grey cat is gone → Le chat gris a disparu*

But

Not OK: *The grey animal is gone → Le gris des animaux est parti (GT, 07/26/09)*

- Long-distance dependencies are often incorrectly handled

Ich habe vorgestern das grüne, komplizierte, von Goethe geschriebene Buch gelesen.
→ I have the green yesterday, complicated, read book written by Goethe (GT, 08/24/09).

- Semantic entities and relations often altered by incorrect ordering

*This might affect the quality of roads, bridges, and **highway finances**.*

*→ Ceci pourrait affecter la qualité des routes, des ponts, **des finances et de l'autoroute**.* (GT, 07/17/2009)

*Marie et Jean **plaisent à ma mère***.

*→ Mary and John **like my mother***. (GT, 07/17/2009)

***John** gave **Mary** a book*.
*→ **Jean-Marie** a donné un livre*. (GT, 07/17/2009)

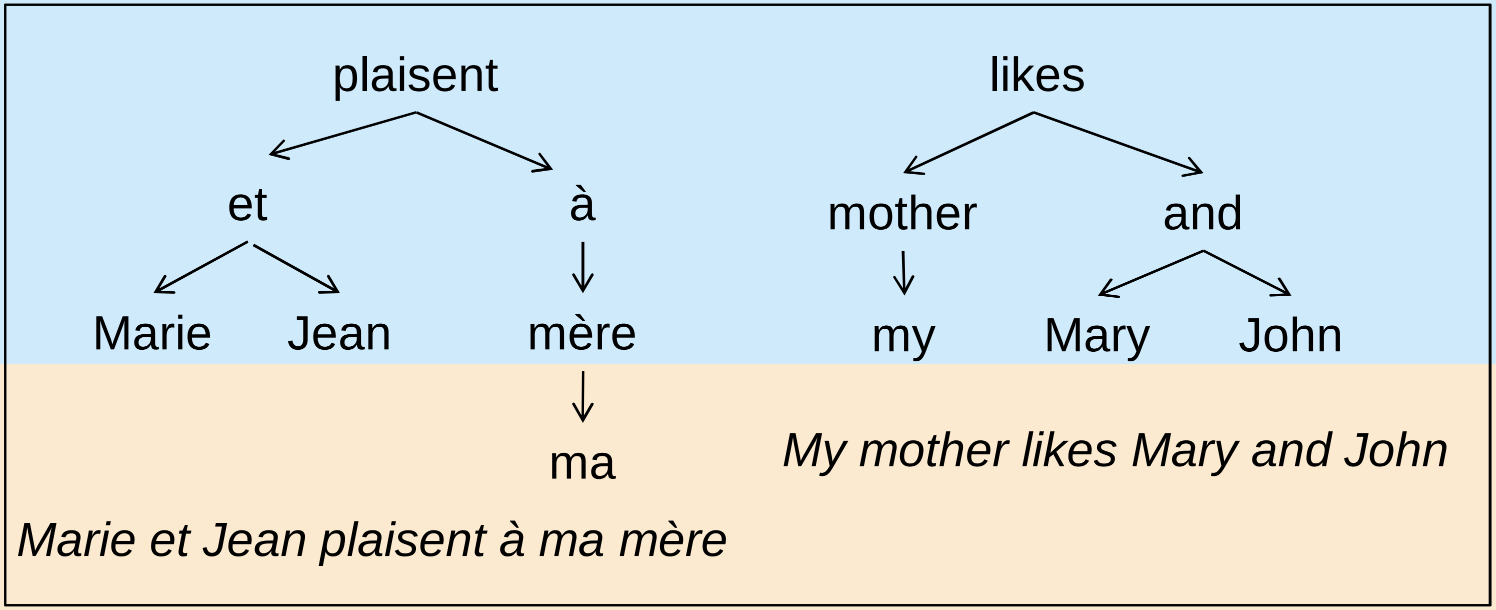
NRC·CNRC
Institute for Information Technology

7. Syntax & reordering in SMT systems (C) String-based Vs tree-based

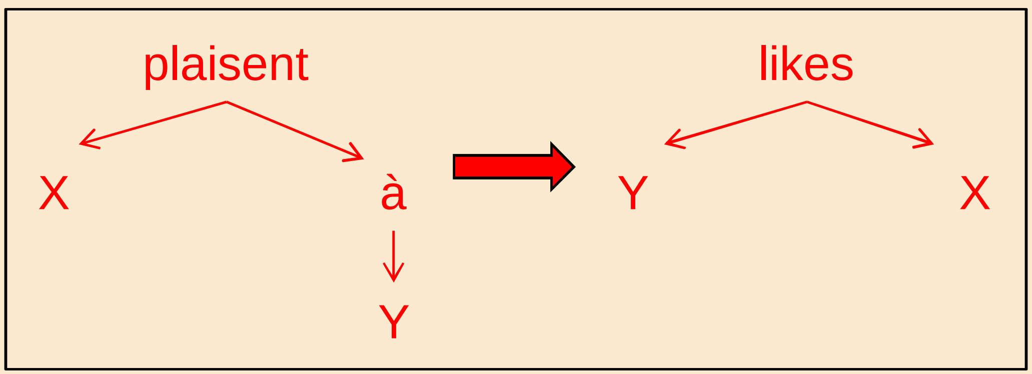
- Classical SMT relies on a *string-based* approach
  - Sentences have a flat structure
  - IBM models: sentence = string of words
  - PBSMT: sentence = string of « phrases » (≠ syntactic phrases)
- But word order phenomena are often difficult to capture at level of strings
- Traditional linguistics relies on syntactic approach: tree-based
  - Sentences possess *tree structure* (hierarchical as opposed to flat)
  - Tree nodes can have grammatical types (*NP*, *PP*, *VP*…)
  - Tree arcs can represent grammatical relations (*subject*, *object*, etc)
  - Ordering rules relative to node types and grammatical relations
- SMT community now moving towards syntax-based models
- Background: Victor Yngve, *A Framework for Syntactic Translation* (1959)

7. Syntax & reordering in SMT systems (C) .. Tree-based: example

Syntactically-annotated bilingual corpus



Learned rules



Where X and Y stand for syntactic phrases of arbitrary size

7. Syntax & reordering in SMT systems (D) Grammar-based SMT models

Varieties of grammar-based models in SMT

- Tree-to-tree models: (Ambati, Lavie, Carbonell)
  - Trained using parsers of both SL and TL plus GIZA word alignment
    - Induce tree correspondence rules
  - Such rules are often cast as synchronous CFG's; thus decoding = synchronous parsing
- String-to-tree models (Galley & al. 2006):
  - Trained using TL parser plus GIZA word alignment
    - Induce string-to-tree transducer
  - Decoding: string-to-tree transduction
- Tree-to-string models (Liu & Gildea, 2008):
  - Trained using SL parser plus GIZA word alignment
    - Induce tree-to-string transducer
  - Decoding: tree-to-string transduction

7. Syntax & reordering in SMT systems (D) Grammar-based models (cont.)

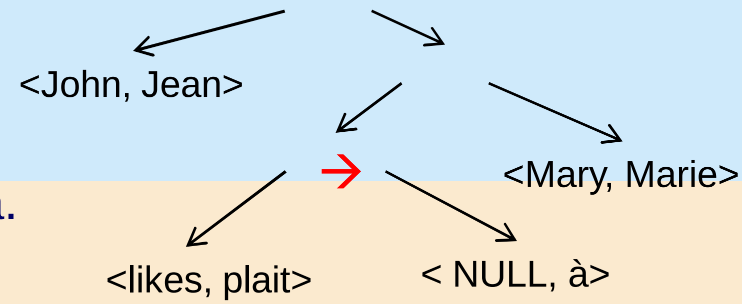
- Early SMT syntactic models had worse results than PBSMT
  - Phrase pairs limited to corresponding complete syntactic units → harmful
  - Only used minimal phrase pairs → lack of context
  - More recent models have at least partly corrected these problems
- Advantages
  - Better overall handling of word order
  - Better at translating discontinuous phrases (E.g. *as X as Y* → *aussi X que Y*)
  - Especially advantageous for handling typologically different languages
  - Fast and steady improvement in recent years: ISI's system obtained best performance on *Ch*→*En* at NIST 2009
- Drawbacks
  - Require expensive language-specific resources (parsers)
  - Performance heavily dependent on parsing quality
  - Larger search space → costlier processing

NRC·CNRC
Institute for Information Technology

7. Syntax & reordering in SMT systems (E) « Formal syntax » models (cont.)

- No linguistic grammar, but induction of hierarchical word/phrase alignment structure from bilingual corpora.

- Wu's inversion transduction grammars (ITG's); unsupervised creation of word-based hierarchical alignment in corpora.

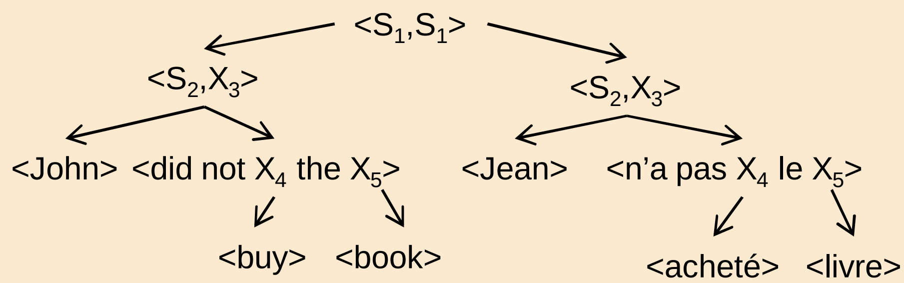


- David Chiang's Hiero MT model:

<did not like the book, n'a pas aimé le livre>
<did not see the cat, n'a pas vu le chat>
…

<did not X_1 the X_2, n'a pas X_1 le X_2>

Generalizing over standard PBSMT phrase pairs



Synchronous CFG parsing with generalized phrases

7. Syntax & reordering in SMT systems (E) « Formal syntax » models (cont.)

- (Xiong, Zhang, Aw and Li) : Enrich Hiero's formal syntax base with some basic linguistic knowledge
- Advantages of formal syntax models:
  - Hierarchical structure makes it possible to account for problems such as:
    - Long-distance dependencies
    - Discontinuous constituents
  - No need for language-specific resources
  - For Ch → En results are better than PBMST and close to those of the best grammar-based models (cf. BBN's system, 2nd at NIST-09).
- Limitations:
  - Lack of grammatical typing often leads to over-generalization

$$\text{give } X_1\ X_2 \neq \textit{give } NP_1\ NP_2$$

7. Syntax & reordering in SMT systems
(F) Enhancing PBSMT models

Approach 1: Syntactic pre-processing (Diaz de Ilarraza, Labaka, Sarosola)

- Use pre-processing component to reorder *SL* → *SL'*
  - Make *SL'* ordering similar to *TL* ordering
  - Handcrafted parse/reorder rules or rules automatically learned from word aligned corpus
- Phrase-based decoder used in « monotonic » mode
- Advantages:
  - Decoding greatly simplified
  - Long-distance dependencies can in principle be tackled (given suitable pre-processing)
- Problems:
  - Serial process: errors from pre-processor difficult to repair downstream
  - Since *SL'* is a pseudo-language, no LM is available to help filter out bad reorderings

7. Syntax & reordering in SMT systems (F) Enhancing PBSMT models (cont.)

Approach 2: Syntactic post-processing

- Rerank *n*-best list of translations produced by decoder
  - Use any kind of syntactic model; assign parsing scores
  - Little success thus far (see Och & al. 2004)
  - Apparently, reasonably-sized *n*-best list does not contain enough variety
- Reordering component at post-processing stage (Na, Li, Kim & Lee 2009)
  - Training: using word alignments, reorder *TL -> TL'* such that TL' order is similar to SL order
  - PBSMT decoder in monotonic mode
  - Post-processing: reorder *TL' → TL;* use a non projective dependency parser

NRC·CNRC
Institute for Information Technology

7. Syntax & reordering in SMT systems (F) Enhancing PBSMT models (cont.)

Approach 3: Phrase-based decoding with syntactic constraints

- Use linguistically-informed parser to guide decoding
  - Penalize decoder paths that yield non-cohesive reordering (Cherry 2008)
  - Formalize reordering as permissible sequences of subtree movements in SL dependency tree (Bach, Gao, Vogel 2009)
- Incorporate « formal parsing » mechanism to PB decoder; decoder combines input phrases into higher-order phrases, and allows movement across these
  - *Chunking* approach (Yahyaei & Monz 2009):
    - Use alignments to learn how chunk SL into a sequence of monotonically translatable groups
  - *Shift-reduce parsing* approach (Galley & Manning 2008)
    - Built-in parser can recursively combine adjacent phrases
- Combine above two approaches (Nguyen, Shimazu & Nguyen 2009)

Conclusions

- More data, faster machines; research on core SMT algorithms flourishing
- SMT is absorbing older approaches (syntax, knowledge in RBMT systems)
- Increasing competition between research teams : SMT getting better & faster
- **But** there are some important gaps:
  - Not enough work on user studies or incorporating MT into translators' tools
  - Good work on more accurate automatic metrics, but these underutilized
  - Too much focus on language pairs where one of the two languages is English; not enough work on morphologically rich languages.
- Let's talk about bridging such gaps during this conference
- This is an exciting time for MT; SMT is generating unprecedented amount of research activity.
- Expectations are high again? Will they be met this time?