Machine Translation: Modeling, Algorithm and Knowledge

Min Zhang

Institute for Infocomm Research, Singapore

20 September 2012

CWMT 2012, Xi'An

MT Sample Results

(S) Did you know that your investment advisor will tell you that fine wine investment consistently delivers between 12 and 30% growth in value? From mid-2009 to 2011, the APM Fine Wine Index, which tracks price movements among investment wines, rose to 176 from 103 points.

(T1) 你知道吗，你的投资顾问会告诉你，始终如一地提供优质葡萄酒投资12至30％的增长值吗？从2009年年中到2011年的APM美酒指数，追踪在投资葡萄酒的价格走势，从103点上升到176。

(T2) 你知道你的投资顾问会告诉你，葡萄酒投资始终提供之间的12和30%的增长，价值？从2009年2011，杀伤人员地雷的精品葡萄酒指数，跟踪价格变动之间的投资葡萄酒，上升至176，103分

What is SMT?

SMT = Linguistic / Any Modeling

+ Statistical decision

Component:

Modeling / Training / Decoding

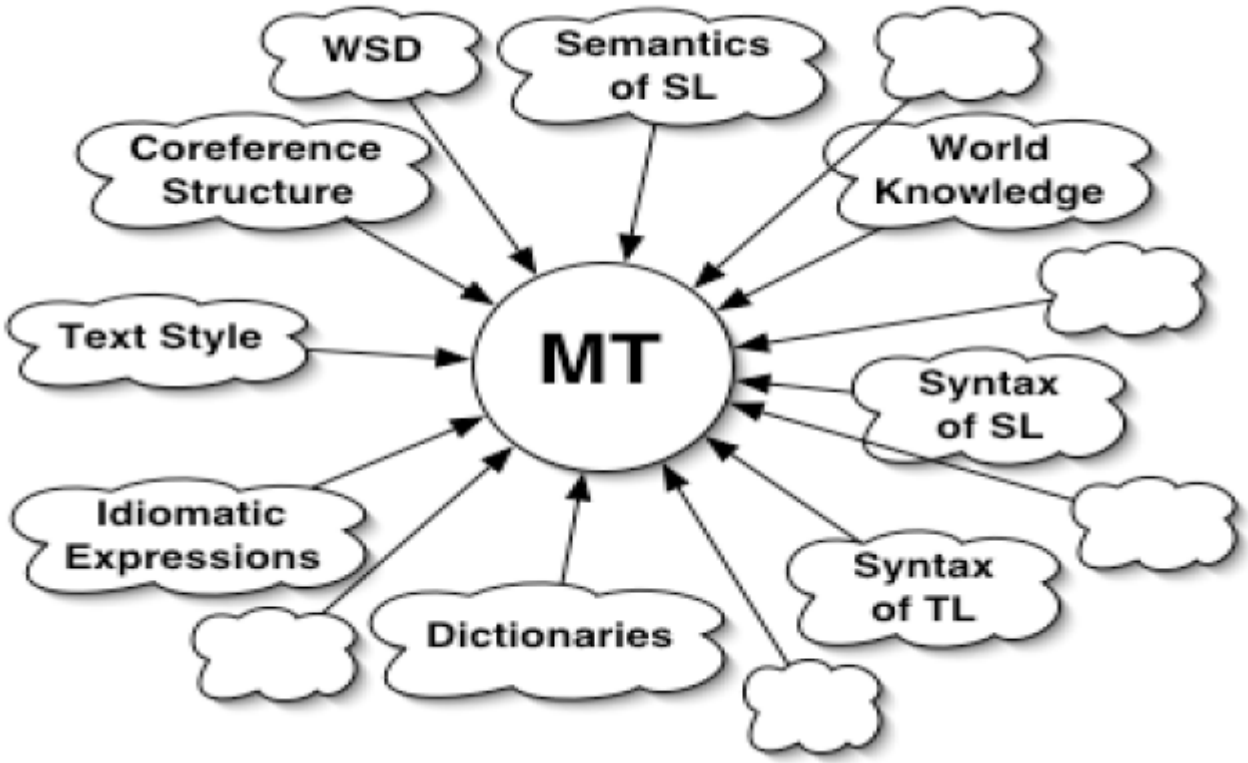
The State-of-the-Art

- Already achieved significant progress in recent 10 years in research and industry, but still in its infant stage.
- Many methodologies proposed and systems deployed, but still have too many issues to be solved. All the issues mixed, need to figure out the most key points.
- Modeling is one of the key points.
- Viewpoints: Research vs. Industry.

Modeling and Models

- Word-based
- Phrase-based based
- Syntax-based
- Shallow semantics-base
- ……

Knowledge Sources (from Och's slide)



MT: An AI-Hard Problem

- Word Order
- Word Sense
- Idioms
- Structure Divergence
- Semantic, Concept, Culture Difference
- MT is A Understanding Issue, the core of AI…

Amazing?

Training	Dev	Test	Bleu
LDC (5M)	NIST02	NIST05 (4 Refs)	0.38
		Part of Training Set (1 Ref)	>0.79

Issues: too many worth exploring

- Over dependency on parallel corpus
- Poor generalization ability
- Structure divergence
- Advanced monolingual NLP technologies do not work well at SMT
- .....

Solutions, Research Topics

- MT-oriented Multilingual Grammar Induction
- Non-parallel-corpora (monolingual) for SMT
- Semantics-based SMT

Machine Translation
Exciting, Enjoyable R&D!

Thanks!