First Year Progress

MOSES  CORE

Hieu Hoang

Luxembourg 2013

Achievements

- Cross-Platform Compatibility
- Ease of use / Installation
- Testing and Reliability
- Speed
- Language Model training with IRSTLM
- Sparse Features
- Making use of TMX
- Lots more by other developers…

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Achievements

- Cross-Platform Compatibility
- Ease of use / Installation
- Testing and Reliability
- Speed
- Language Model training with IRSTLM
- Sparse Features
- Making use of TMX
- Lots more by other developers…

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Achievements

- Cross-Platform Compatibility
- Ease of use / Installation
- Testing and Reliability
- Speed
- Language Model training with IRSTLM
- Sparse Features
- Making use of TMX
- Lots more by other developers…

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Achievements

- Cross-Platform Compatibility
- Ease of use / Installation
- Testing and Reliability
- Speed
- Language Model training with IRSTLM
- Sparse Features
- Making use of TMX
- Lots more by other developers…

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Cross-Platform Compatibility

- Tested on
  - Windows 7 (32-bit) with Cygwin 6.1
  - Mac OSX 10.7 with MacPorts
  - Ubuntu 12.10, 32 and 64-bit
  - Debian 6.0, 32 and 64-bit
  - Fedora 17, 32 and 64-bit
  - openSUSE 12.2, 32 and 64-bit
- Project files for
  - Visual Studio
  - Eclipse on Linux and Mac OSX

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Ease of use / Installation

- Easier compile and install
  - Boost bjam
  - No installation required
- Binaries available for
  - Linux
  - Mac
  - Windows/Cygwin
  - Moses + Friends
    - IRSTLM
    - GIZA++ and MGIZA
- Ready-made models trained on Europarl (and others)
  - 10 language pairs
  - phrase-based, hierarchical, factored models

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Testing and Reliability

- Monitor check-ins
- Unit tests
- More regression tests
- Nightly tests
  - Run end-to-end training
  - http://www.statmt.org/moses/cruise/
- Tested on all major OSes
- Train Europarl models
  - Phrase-based, hierarchical, factored
  - 8 language-pairs
  - http://www.statmt.org/moses/RELEASE-1.0/models/

Speed
Training

- Multithreaded

Time (mins)	1-core	2-cores	4-cores	8-cores	Size (MB)
Phrase-based	60	47 (79%)	37 (63%)	33 (56%)	893
Hierarchical	1030	677 (65%)	473 (45%)	375 (36%)	8300

- Reduced disk IO
  - compress intermediate files
- Reduce disk space requirement

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Speed
Training

- Multithreaded

Time (mins)	1-core	2-cores	4-cores	8-cores	Size (MB)
Phrase-based	60	47 (79%)	37 (63%)	33 (56%)	893
Hierarchical	1030	677 (65%)	473 (45%)	375 (36%)	8300

- Reduced disk IO
  - compress intermediate files
- Reduce disk space requirement

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Speed
Training

- Multithreaded

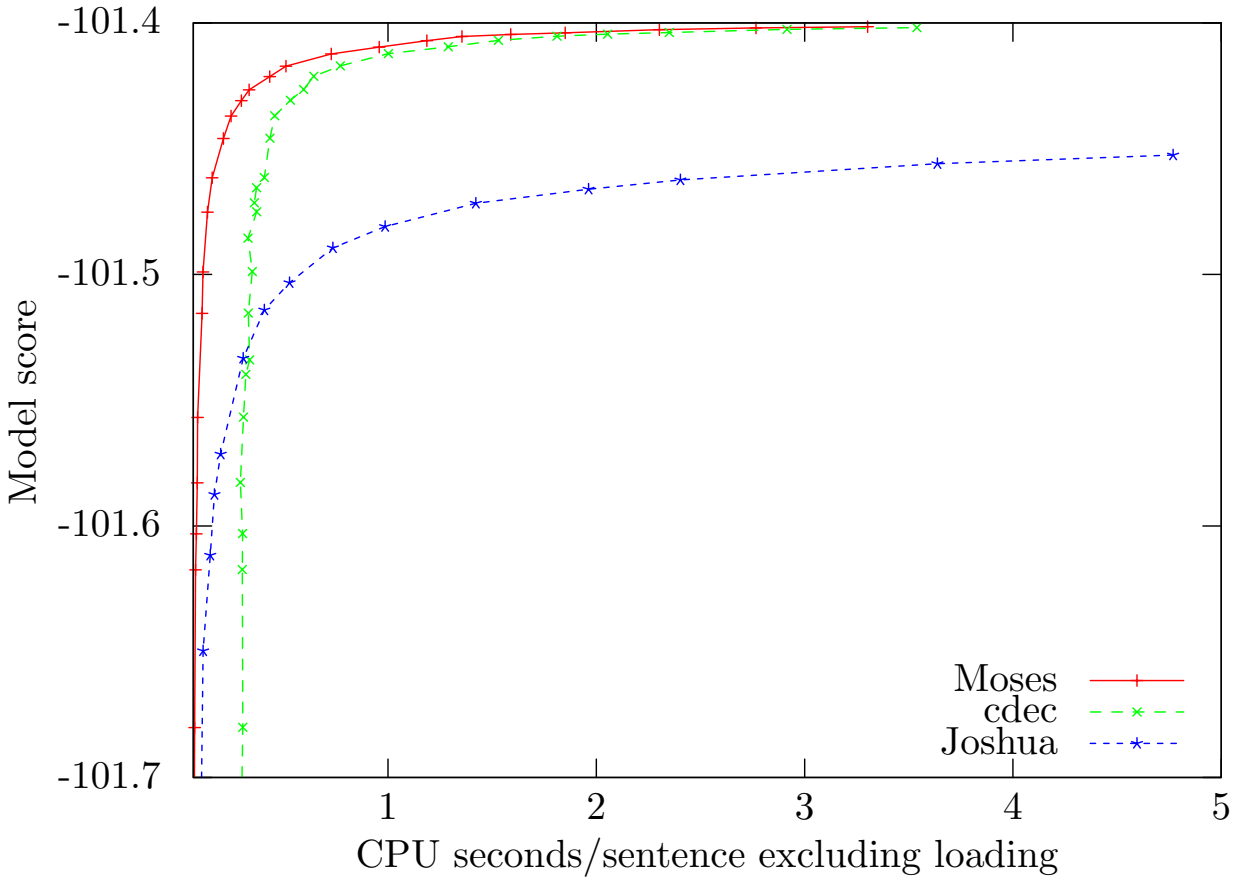
Time (mins)	1-core	2-cores	4-cores	8-cores	Size (MB)
Phrase-based	60	47 (79%)	37 (63%)	33 (56%)	893
Hierarchical	1030	677 (65%)	473 (45%)	375 (36%)	8300

- Reduced disk IO
  - compress intermediate files
- Reduce disk space requirement

MOSES CORE

Speed

Decoding



thanks to Ken!!

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Language Model training with IRSTLM

- IRSTLM v. SRILM
  - Faster
    - Parallelization
  - Train larger language models
    - Uses less memory
  - Open-Source
- Integrated into Moses training pipeline
  - every part of pipeline now Open-Source
- Competition for LM training
  - LM training with KenLM

Sparse Features

- Large number of sparse features
  - 1+ millions
  - Sparse AND dense features
- Available sparse features

Target Bigram	Target Ngram	Source Word Deletion
Sparse Phrase Table	Phrase Boundary	Phrase Length
Phrase Pair	Target Word Insertion	Global Lexical Model

- Different tuning
  - MERT
  - Mira
  - Batch Mira (Cherry & Foster, 2012)
  - PRO (Hopkins and May, 2011)

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Making use of TMX

- Translation Memories
  - created by translators
  - highly repetitive
- In-domain translation
- Better use of TM
  - 'Convergence of Translation Memory and Statistical Machine Translation' AMTA 2010
  - Use TM as templates
  - Preferred over MT rules

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Achievements

- Cross-Platform Compatibility
  - Windows, Mac OSX, Linux
- Ease of use / Installation
  - No installation, binaries for download
- Testing and Reliability
  - Unit tests & regression test, nightly tests, end-to-end experiments
- Sparse Features
  - Millions of features, MIRA & PRO tuning
- Speed
  - Faster training, fastest open-source syntax decoder
- Language Model training with IRSTLM
  - Faster, bigger, open-source
- Making use of TMX
  - Use TMX as templates

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Year 2 Priorities

- Code cleanup
- Incremental Training
- Better translation
  - smaller model
  - bigger data
  - faster training and decoding

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Code cleanup

- Framework for feature functions
  - Easier to add new feature functions

- Cleanup
  - Refactor
  - Delete old code
  - Documentation

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Incremental Training

- Incremental word alignment
- Dynamic suffix array
- Phrase-table update

- Better integration with rest of Moses

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Smaller files

- Smaller binary
  - phrase-tables
  - language models
- Mobile devices
- Fits into memory
  →faster decoding

- Efficient data structures
  - suffix arrays
  - compressed file formats

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Better Translations

- Consistently beat phrase-based models for every language pair

	Phrase-Based	Hierarchical
en-es	24.81	24.20
es-en	23.01	22.37
en-cs	11.04	10.93
cs-en	15.72	15.68
en-de	11.87	11.62
de-en	15.75	15.53
en-fr	22.84	22.28
fr-en	25.08	24.37

MOSES CORE

THE UNIVERSITY OF EDINBURGH

Questions…

MOSES CORE