MaTrEx: Machine Translation Using Examples

Stephen Armstrong, Marian Flanagan, Yvette Graham, Declan Groves, Bart Mellebeek, Sara Morrissey, Nicolas Stroppa and Andy Way

NCLT, School of Computing, Dublin City University

DCU NCLT @ OpenLab2006

Outline

Introduction

- Large-scale Example-Based Machine Translation system
  - Robust
  - Easily adaptable to new language pairs
  - Modular design - follow established Design Patterns
- Built by a team of researchers at the National Centre for Language Technology (NCLT) in DCU
  - 6 Ph.D. Students, 1 Postdoc
  - Supervised by Dr. Andy Way
- First participation of an EBMT system in a shared task

Example-Based MT

- Based on the intuition that humans make use of previously seen translation examples to translate unseen input
  - Analogy-based principle

Example-Based MT

- Based on the intuition that humans make use of previously seen translation examples to translate unseen input
  - Analogy-based principle
- As with SMT, makes use of information extracted from sententially-aligned corpora

Example-Based MT

- Based on the intuition that humans make use of previously seen translation examples to translate unseen input
    - Analogy-based principle
- As with SMT, makes use of information extracted from sententially-aligned corpora
- Translation performed using database of examples extracted from corpora

Example-Based MT

- Based on the intuition that humans make use of previously seen translation examples to translate unseen input
    - Analogy-based principle
- As with SMT, makes use of information extracted from sententially-aligned corpora
- Translation performed using database of examples extracted from corpora
- During translation, the input sentence is matched against the example database and corresponding target language examples are recombined to produce final translation.

EBMT: *An Example*

- Assume an aligned bilingual corpus of examples against which input text is matched
- Best match is found using a similarity metric (can be based on word co-occurrence, POS, bilingual dictionaries etc.)

Given the Corpus

La tienda abrió el lunes pasado = The shop opened last Monday
Juan fue a la piscina = John went to the swimming pool
La carnicerna está al lado de la panadería = The butcher's is next to the baker's

EBMT: *An Example*

- Identify useful fragments

Given the Corpus

La tienda abrió el **lunes pasado** = *The shop opened* **last Monday**
Juan fue a *la piscina* = **John went to** *the swimming pool*
La carnicerna está al lado de **la panadería** = *The butcher's is next to* **the baker's**

EBMT: *An Example*

- Identify useful fragments
- Recombine extracted fragments to translate new unseen input

Given the Corpus

La tienda abrió el **lunes pasado** = *The shop opened* **last Monday**
Juan fue a *la piscina* = **John went to** *the swimming pool*
La carnicerna está al lado de **la panadería** = *The butcher's is next to* **the baker's**

Translate New Input

Juan fue a la panadería el lunes pasado = John went to the baker's last Monday

Marker-Based EBMT

- Approach to EBMT based on the Marker Hypothesis

> *"The Marker Hypothesis states that all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context." (Green, 1979).*

- Universal psycholinguistic constraint: languages are marked for syntactic structure at surface level by closed set of lexemes or morphemes.

Marker-Based EBMT

- Approach to EBMT based on the Marker Hypothesis

> *"The Marker Hypothesis states that all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context." (Green, 1979).*

- Universal psycholinguistic constraint: languages are marked for syntactic structure at surface level by closed set of lexemes or morphemes.

The Dearborn Mich., energy company stopped paying a dividend in the third quarter of 1984 because of troubles at its Midland nuclear plant

Marker-Based EBMT

- Approach to EBMT based on the Marker Hypothesis

> *"The Marker Hypothesis states that all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context." (Green, 1979).*

- Universal psycholinguistic constraint: languages are marked for syntactic structure at surface level by closed set of lexemes or morphemes.

The Dearborn Mich., *energy company stopped paying* **a dividend** *in* **the third quarter** *of 1984 because of troubles at* **its Midland nuclear plant**

- 3 NPs start with determiners, one with a possessive pronoun
  - Determiners & possessive pronoun - small closed-class sets
  - Predicts head nominal element will occur in the right-context.

Marker-Based EBMT

- Approach to EBMT based on the Marker Hypothesis

  "The Marker Hypothesis states that all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context." (Green, 1979).

- Universal psycholinguistic constraint: languages are marked for syntactic structure at surface level by closed set of lexemes or morphemes.

The Dearborn Mich., energy company stopped paying a dividend ***in*** *the third quarter* ***of*** *1984 because* ***of*** *troubles* ***at*** *its Midland nuclear plant*

- 3 NPs start with determiners, one with a possessive pronoun
  - Determiners & possessive pronoun - small closed-class sets
  - Predicts head nominal element will occur in the right-context.
- Four prepositional phrases, with prepositional heads.
  - Again a small set of closed-class words
  - Indicates that soon thereafter an NP object will occur

Marker-Based EBMT: *Previous Work*

- Line of previous research:
  - (Gough et al., 2002) *AMTA*
  - (Gough & Way, 2003) *MT Summit*
  - (Way & Gough, 2003) *Computational Linguistics*
  - (Gough & Way, 2004) *EAMT*
  - (Way & Gough, 2004) *TMI*
  - (Gough, 2005) *Ph.D. Thesis*
  - (Way & Gough, 2005) *Natural Language Engineering*
  - (Way & Gough, 2005) *Machine Translation*
  - (Groves & Way, 2004) *ACL Workshop on Data-Driven MT*
  - (Groves & Way, 2005) *MT Journal Special Issue on EBMT*

Marker-Based EBMT: *Previous Work*

- Line of previous research:
  - (Gough et al., 2002) *AMTA*
  - (Gough & Way, 2003) *MT Summit*
  - (Way & Gough, 2003) *Computational Linguistics*
  - (Gough & Way, 2004) *EAMT*
  - (Way & Gough, 2004) *TMI*
  - (Gough, 2005) *Ph.D. Thesis*
  - (Way & Gough, 2005) *Natural Language Engineering*
  - (Way & Gough, 2005) *Machine Translation*
  - (Groves & Way, 2004) *ACL Workshop on Data-Driven MT*
  - (Groves & Way, 2005) *MT Journal Special Issue on EBMT*
- Have previously only worked on French-English and German-English data
- Largest training data set used to date consisted of 322K French-English sentence pairs

Marker-Based EBMT: *Previous Work*

- Line of previous research:
  - (Gough et al., 2002) *AMTA*
  - (Gough & Way, 2003) *MT Summit*
  - (Way & Gough, 2003) *Computational Linguistics*
  - (Gough & Way, 2004) *EAMT*
  - (Way & Gough, 2004) *TMI*
  - (Gough, 2005) *Ph.D. Thesis*
  - (Way & Gough, 2005) *Natural Language Engineering*
  - (Way & Gough, 2005) *Machine Translation*
  - (Groves & Way, 2004) *ACL Workshop on Data-Driven MT*
  - (Groves & Way, 2005) *MT Journal Special Issue on EBMT*
- Have previously only worked on French-English and German-English data
- Largest training data set used to date consisted of 322K French-English sentence pairs
- MaTrEx system is a complete re-implementation of previous system
  - More sophisticated marker sets and marker-based chunk alignment

Marker-Based EBMT: *Chunking*

- Use a set of closed-class marker words to segment aligned source and target sentences during a pre-processing stage.
- <PUNC> used as end of chunk marker

Determiner	<DET>
Quantifiers	<Q>
Prepositions	<P>
Conjunctions	<C>
WH-Adverbs	<WH>
Possessive Pronouns	<POSS-PRON>
Personal Pronouns	<PERS-PRON>
Punctuation Marks	<PUNC>

Marker-Based EBMT: *Chunking*

- Use a set of closed-class marker words to segment aligned source and target sentences during a pre-processing stage.
- <PUNC> used as end of chunk marker

Determiner	<DET>
Quantifiers	<Q>
Prepositions	<P>
Conjunctions	<C>
WH-Adverbs	<WH>
Possessive Pronouns	<POSS-PRON>
Personal Pronouns	<PERS-PRON>
Punctuation Marks	<PUNC>

- English Marker words extracted from CELEX and edited manually to correspond with the training data.
- Spanish Marker words from 2 stop word lists, generously supplied by Lluís Padró (Polytechnic University of Catalunya) and Montserrat Civit (University of Barcelona).

Marker-Based EBMT: *Chunking (2)*

- Enables the use of basic syntactic marking for extraction of translation resources
- Source-target sentence pairs are tagged with their marker categories automatically in a pre-processing step:

SP: *<PRON> Usted cliquea <PREP> en <DET> el botón rojo <PREP> para ver <DET> el efecto <PREP> de <DET> la selección.*

EN: *<PRON> You click <PREP> on <DET> the red button <PREP> to view <DET> the effect <PREP> of <DET>the selection*

Marker-Based EBMT: *Chunking (2)*

- Enables the use of basic syntactic marking for extraction of translation resources
- Source-target sentence pairs are tagged with their marker categories automatically in a pre-processing step:

SP: *<PRON> Usted cliquea <PREP> en <DET> el botón rojo <PREP> para ver <DET> el efecto <PREP> de <DET> la selección.*

EN: *<PRON> You click <PREP> on <DET> the red button <PREP> to view <DET> the effect <PREP> of <DET>the selection*

- Aligned source-target chunks are created by segmenting the sentence based on these tags, along with word translation probability and cognate information:

<PRON> Usted cliquea	:	*<PRON> You click*
<PREP> en el botón rojo	:	*<PREP> on the red button*
<PREP> para ver	:	*<PREP> to view*
<DET> el efecto	:	*<DET> the effect*
<PREP> de la selección	:	*<PREP> of the selection*

Marker-Based EBMT: *Chunking (2)*

- Enables the use of basic syntactic marking for extraction of translation resources
- Source-target sentence pairs are tagged with their marker categories automatically in a pre-processing step:

SP: *<PRON> Usted cliquea* **<PREP> en <DET> el botón rojo** *<PREP> para ver <DET> el efecto <PREP> de <DET> la selección.*

EN: *<PRON> You click* **<PREP> on <DET> the red button** *<PREP> to view <DET> the effect <PREP> of <DET>the selection*

- Aligned source-target chunks are created by segmenting the sentence based on these tags, along with word translation probability and cognate information:

<PRON> Usted cliquea	:	*<PRON> You click*
<PREP> en el botón rojo	:	**<PREP> on the red button**
<PREP> para ver	:	*<PREP> to view*
<DET> el efecto	:	*<DET> the effect*
<PREP> de la selección	:	*<PREP> of the selection*

- Chunks must contain at least one non-marker word - ensures chunks contain useful contextual information

Chunk Alignment

- Focused on chunk alignment for this task
  - Discriminative Approach for chunk alignment
- "Edit-Distance" Chunk Alignment
  - Dynamic programming

Chunk Alignment

- Focused on chunk alignment for this task
    - Discriminative Approach for chunk alignment
- "Edit-Distance" Chunk Alignment
    - Dynamic programming
- Distance metrics used:
    - Distance based on Marker Tags

Chunk Alignment

- Focused on chunk alignment for this task
  - Discriminative Approach for chunk alignment
- "Edit-Distance" Chunk Alignment
  - Dynamic programming
- Distance metrics used:
  - Distance based on Marker Tags
  - Chunk Minimum Edit-Distance: *Word-Based Distance, Character-Based Distance*

Chunk Alignment

- Focused on chunk alignment for this task
  - Discriminative Approach for chunk alignment
- "Edit-Distance" Chunk Alignment
  - Dynamic programming
- Distance metrics used:
  - Distance based on Marker Tags
  - Chunk Minimum Edit-Distance: *Word-Based Distance, Character-Based Distance*
  - Cognate Information: *Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance*

Chunk Alignment

- Focused on chunk alignment for this task
  - Discriminative Approach for chunk alignment
- "Edit-Distance" Chunk Alignment
  - Dynamic programming
- Distance metrics used:
  - Distance based on Marker Tags
  - Chunk Minimum Edit-Distance: *Word-Based Distance, Character-Based Distance*
  - Cognate Information: *Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance*
  - Word Translation Probabilities

Chunk Alignment

- Focused on chunk alignment for this task
  - Discriminative Approach for chunk alignment
- "Edit-Distance" Chunk Alignment
  - Dynamic programming
- Distance metrics used:
  - Distance based on Marker Tags
  - Chunk Minimum Edit-Distance: *Word-Based Distance, Character-Based Distance*
  - Cognate Information: *Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance*
  - Word Translation Probabilities
  - Combination (can be viewed as a log-linear model)

$$\lambda_1 d_1(a|b) + ...\lambda_n d_n(a|b) \Rightarrow -\lambda_1 logP_1(a|b)... - \lambda_n logP_n(a|b)$$

Chunk Alignment

- Focused on chunk alignment for this task
  - Discriminative Approach for chunk alignment
- "Edit-Distance" Chunk Alignment
  - Dynamic programming
- Distance metrics used:
  - Distance based on Marker Tags
  - Chunk Minimum Edit-Distance: *Word-Based Distance, Character-Based Distance*
  - Cognate Information: *Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance*
  - Word Translation Probabilities
  - Combination (can be viewed as a log-linear model)

$$\lambda_1 d_1(a|b) + \ldots \lambda_n d_n(a|b) \Rightarrow -\lambda_1 logP_1(a|b) \ldots - \lambda_n logP_n(a|b)$$

- "Edit-Distance" with Jumps
  - Found that this method did not improve results - similar chunk order between Spanish and English

Word Alignment

- "Refined" method of (Och & Ney, 2003)

Word Alignment

- "Refined" method of (Och & Ney, 2003)
  - Use GIZA++ word alignment tool to perform Spanish-English and English-Spanish word alignment

Word Alignment

- "Refined" method of (Och & Ney, 2003)
  - Use GIZA++ word alignment tool to perform Spanish-English and English-Spanish word alignment
  - Take the intersection of these uni-directional alignment sets - gives a set of highly confident alignments

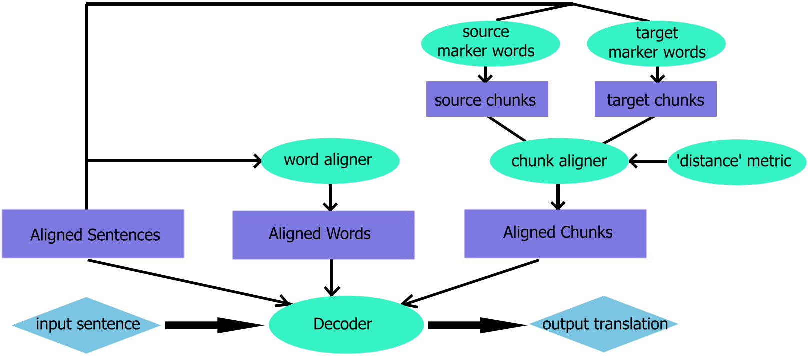
Word Alignment

- "Refined" method of (Och & Ney, 2003)
  - Use GIZA++ word alignment tool to perform Spanish-English and English-Spanish word alignment
  - Take the intersection of these uni-directional alignment sets - gives a set of highly confident alignments
  - Extend this intersection into the union of the alignment sets, by iteratively adding adjacent alignments

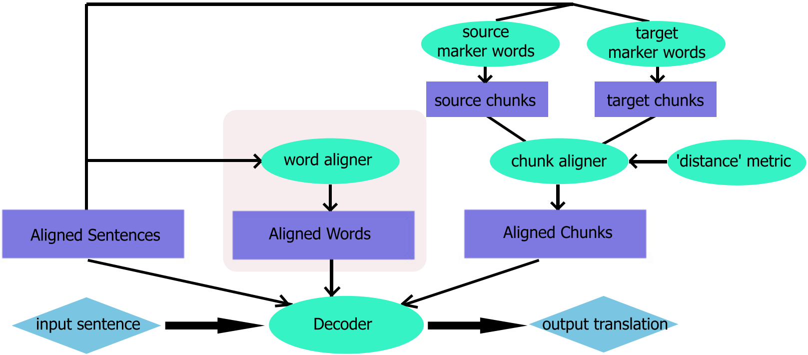
Word Alignment

- "Refined" method of (Och & Ney, 2003)
  - Use GIZA++ word alignment tool to perform Spanish-English and English-Spanish word alignment
  - Take the intersection of these uni-directional alignment sets - gives a set of highly confident alignments
  - Extend this intersection into the union of the alignment sets, by iteratively adding adjacent alignments
- Only made use of the resulting one-to-one word alignments produced
- Word probabilities were then estimated from relative frequencies.

System Architecture

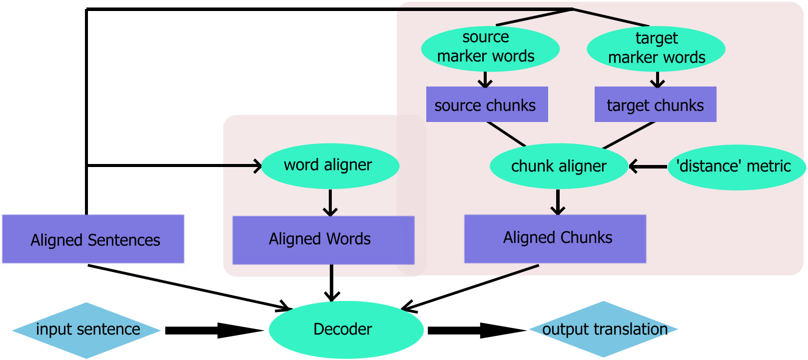


System Architecture



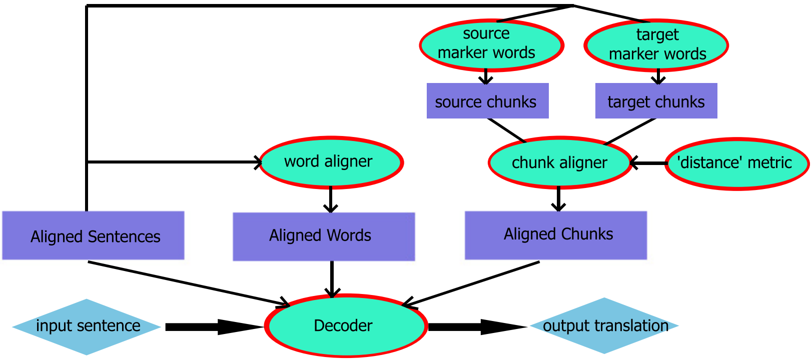
- Aligned Sentences are submitted to word alignment and chunk alignment modules to produce translation resources
- Modular in design
- Easily adaptable and extendible

System Architecture



- Aligned Sentences are submitted to word alignment and chunk alignment modules to produce translation resources
- Modular in design
- Easily adaptable and extendible

System Architecture



- Aligned Sentences are submitted to word alignment and chunk alignment modules to produce translation resources
- Modular in design
- Easily adaptable and extendible
- Modules can be replaced by different implementations

Experiments and Results

- Data used:
  - Filtered supplied Spanish-English training data based on sentence length ($<$ 40 words) and relative sentence length ratio (1.5).
    - 23.4% filtered based on length, 1.8% filtered based on ratio.
  - Text was lowercased
  - Resulted in approx 958K sentence pairs which were used for training.
  - Training took approx. 3hrs on 64-bit machine with 8GB RAM. Testing took 30mins approx.

Experiments and Results

- Data used:
  - Filtered supplied Spanish-English training data based on sentence length ($<$ 40 words) and relative sentence length ratio (1.5).
    - 23.4% filtered based on length, 1.8% filtered based on ratio.
  - Text was lowercased
  - Resulted in approx 958K sentence pairs which were used for training.
  - Training took approx. 3hrs on 64-bit machine with 8GB RAM. Testing took 30mins approx.
- Performed Spanish-English translation.
  - Pharaoh Phrase-Based Decoder (Koehn, 2004)
  - Edit-Distance Chunk Alignment
  - Various combinations of distance metrics weighted linearly

Experiments and Results

- Data used:
  - Filtered supplied Spanish-English training data based on sentence length ($<$ 40 words) and relative sentence length ratio (1.5).
    - 23.4% filtered based on length, 1.8% filtered based on ratio.
  - Text was lowercased
  - Resulted in approx 958K sentence pairs which were used for training.
  - Training took approx. 3hrs on 64-bit machine with 8GB RAM. Testing took 30mins approx.
- Performed Spanish-English translation.
  - Pharaoh Phrase-Based Decoder (Koehn, 2004)
  - Edit-Distance Chunk Alignment
  - Various combinations of distance metrics weighted linearly
- Baseline system: "refined" word alignments passed to Pharaoh decoder.

Results

	BLEU	NIST	CER	PER	WER
Baseline	0.3630	8.3237	51.6662	34.6757	60.2711
Cog,Tag	0.4039	8.7712	44.8441	33.3748	53.2294
WordP,Tag	0.4077	8.8294	44.8192	33.1391	53.3386
Cog,WordP,Tag	0.4092	8.8498	44.6697	33.0518	53.1159

- Baseline achieves high performance only using word information.
  - How often are phrases actually used by Pharaoh?

Results

	BLEU	NIST	CER	PER	WER
Baseline	0.3630	8.3237	51.6662	34.6757	60.2711
Cog,Tag	0.4039	8.7712	44.8441	33.3748	53.2294
WordP,Tag	0.4077	8.8294	44.8192	33.1391	53.3386
Cog,WordP,Tag	0.4092	8.8498	44.6697	33.0518	53.1159

- Baseline achieves high performance only using word information.
    - How often are phrases actually used by Pharaoh?
- Best performing distance metric uses cognate information, word probabilities and marker tags
- We get a relative increase of 12.31% BLEU score over the baseline (during development a max. BLEU score of 0.42 was achieved)

Results

	BLEU	NIST	CER	PER	WER
Baseline	0.3630	8.3237	51.6662	34.6757	60.2711
Cog,Tag	0.4039	8.7712	44.8441	33.3748	53.2294
WordP,Tag	0.4077	8.8294	44.8192	33.1391	53.3386
Cog,WordP,Tag	0.4092	8.8498	44.6697	33.0518	53.1159

- Baseline achieves high performance only using word information.
    - How often are phrases actually used by Pharaoh?
- Best performing distance metric uses cognate information, word probabilities and marker tags
- We get a relative increase of 12.31% BLEU score over the baseline (during development a max. BLEU score of 0.42 was achieved)
- However, should compare system against baseline phrase-based system

Results: *Sample Translations*

- The MaTrEx system often uses good turn of phrase during translation and produces much more coherent output

Results: *Sample Translations*

- The MaTrEx system often uses good turn of phrase during translation and produces much more coherent output

Baseline: *the report that we, the european union and equipping of 21,000 million euros to saudi arabia*

MaTrEx: *the report we are discussing the european union cashed arms and military equipment to the tune of millions of euro in countries such as saudi arabia*

Ref: *in the report we are currently discussing, the european union sold arms and military equipment to the value of 21 billion euros in countries such as saudi arabia*

Results: *Sample Translations*

- The MaTrEx system often uses good turn of phrase during translation and produces much more coherent output

Baseline: *the report that we, the european union* **and equipping of 21,000 million euros** *to saudi arabia*

MaTrEx: *the report we are discussing the european union* **cashed arms and military equipment to the tune of millions of euro** *in countries such as saudi arabia*

Ref: *in the report we are currently discussing, the european union* **sold arms and military equipment to the value of 21 billion euros** *in countries such as saudi arabia*

Results: *Sample Translations*

- The MaTrEx system often uses good turn of phrase during translation and produces much more coherent output

Baseline: *the report that we, the european union and equipping of 21,000 million euros to saudi arabia*

MaTrEx: *the report we are discussing the european union cashed arms and military equipment to the tune of millions of euro in countries such as saudi arabia*

Ref: *in the report we are currently discussing, the european union sold arms and military equipment to the value of 21 billion euros in countries such as saudi arabia*

- The use of chunks gives the system enough context to accurately translate noun phrases

Results: *Sample Translations*

- The MaTrEx system often uses good turn of phrase during translation and produces much more coherent output

Baseline: *the report that we, the european union and equipping of 21,000 million euros to saudi arabia*

MaTrEx: *the report we are discussing the european union cashed arms and military equipment to the tune of millions of euro in countries such as saudi arabia*

Ref: *in the report we are currently discussing, the european union sold arms and military equipment to the value of 21 billion euros in countries such as saudi arabia*

- The use of chunks gives the system enough context to accurately translate noun phrases

Baseline: *those countries are convinced that need to cooperate more effectively in the fight against the terrorism. underneath by way of*

MaTrEx: *the netherlands are convinced that we have to work together more effectively in fighting terrorism*

Ref: *the netherlands is convinced that we must cooperate much more efficiently in the fight against terrorism*

Results: *Sample Translations*

- The MaTrEx system often uses good turn of phrase during translation and produces much more coherent output

Baseline: *the report that we, the european union and equipping of 21,000 million euros to saudi arabia*

MaTrEx: *the report we are discussing the european union cashed arms and military equipment to the tune of millions of euro in countries such as saudi arabia*

Ref: *in the report we are currently discussing, the european union sold arms and military equipment to the value of 21 billion euros in countries such as saudi arabia*

- The use of chunks gives the system enough context to accurately translate noun phrases

Baseline: *those countries are convinced that need to cooperate more effectively* **in the fight against the terrorism**. *underneath by way of*

MaTrEx: *the netherlands are convinced that we have to work together more effectively* **in fighting terrorism**

Ref: *the netherlands is convinced that we must cooperate much more efficiently* **in the fight against terrorism**

Discussions and Conclusions

- Introduced the MaTrEx Data-Driven MT system being developed at the NCLT in Dublin City University
  - Modular design - easily adaptable and extendible

Discussions and Conclusions

- Introduced the MaTrEx Data-Driven MT system being developed at the NCLT in Dublin City University
  - Modular design - easily adaptable and extendible
- Marker-based approach for chunking
- Investigated a number of strategies for chunk alignment
  - Aligning based on marker tags, cognate information and word probabilities most effective
  - Using cognate information as accurate as word probabilities

Discussions and Conclusions

- Introduced the MaTrEx Data-Driven MT system being developed at the NCLT in Dublin City University
  - Modular design - easily adaptable and extendible
- Marker-based approach for chunking
- Investigated a number of strategies for chunk alignment
  - Aligning based on marker tags, cognate information and word probabilities most effective
  - Using cognate information as accurate as word probabilities
- System achieves a BLEU score of 0.4092 - a 12.31% relative increase over a word-based baseline system
- Results indicate the high quality of the chunk alignments extracted

Ongoing and Future Work

- Plan to continue the development the MaTrEx system.
  - Currently at early stage of development
- Implement an example-based decoder.
- Implement an HMM chunk alignment strategy.
- Use of generalised templates - allow more flexibility to the matching process, improves coverage and quality

Ongoing and Future Work

- Plan to continue the development the MaTrEx system.
  - Currently at early stage of development
- Implement an example-based decoder.
- Implement an HMM chunk alignment strategy.
- Use of generalised templates - allow more flexibility to the matching process, improves coverage and quality
- Experiment using different data sets and language pairs
  - OpenLab noisy data set
  - Participate in HLT-NAACL: French-English, German-English translation
  - Other bake-offs: NIST, IWSLT...
  - Basque translation

Ongoing and Future Work

- Plan to continue the development the MaTrEx system.
  - Currently at early stage of development
- Implement an example-based decoder.
- Implement an HMM chunk alignment strategy.
- Use of generalised templates - allow more flexibility to the matching process, improves coverage and quality
- Experiment using different data sets and language pairs
  - OpenLab noisy data set
  - Participate in HLT-NAACL: French-English, German-English translation
  - Other bake-offs: NIST, IWSLT...
  - Basque translation
- Use the system for related research:
  - Sign-Language translation
  - Hybrid Models of EBMT and SMT

Thank you

Thank you for your attention.

http://www.computing.dcu.ie/research/nclt