Empirical Machine Translation and its Evaluation

EAMT Best Thesis Award 2008

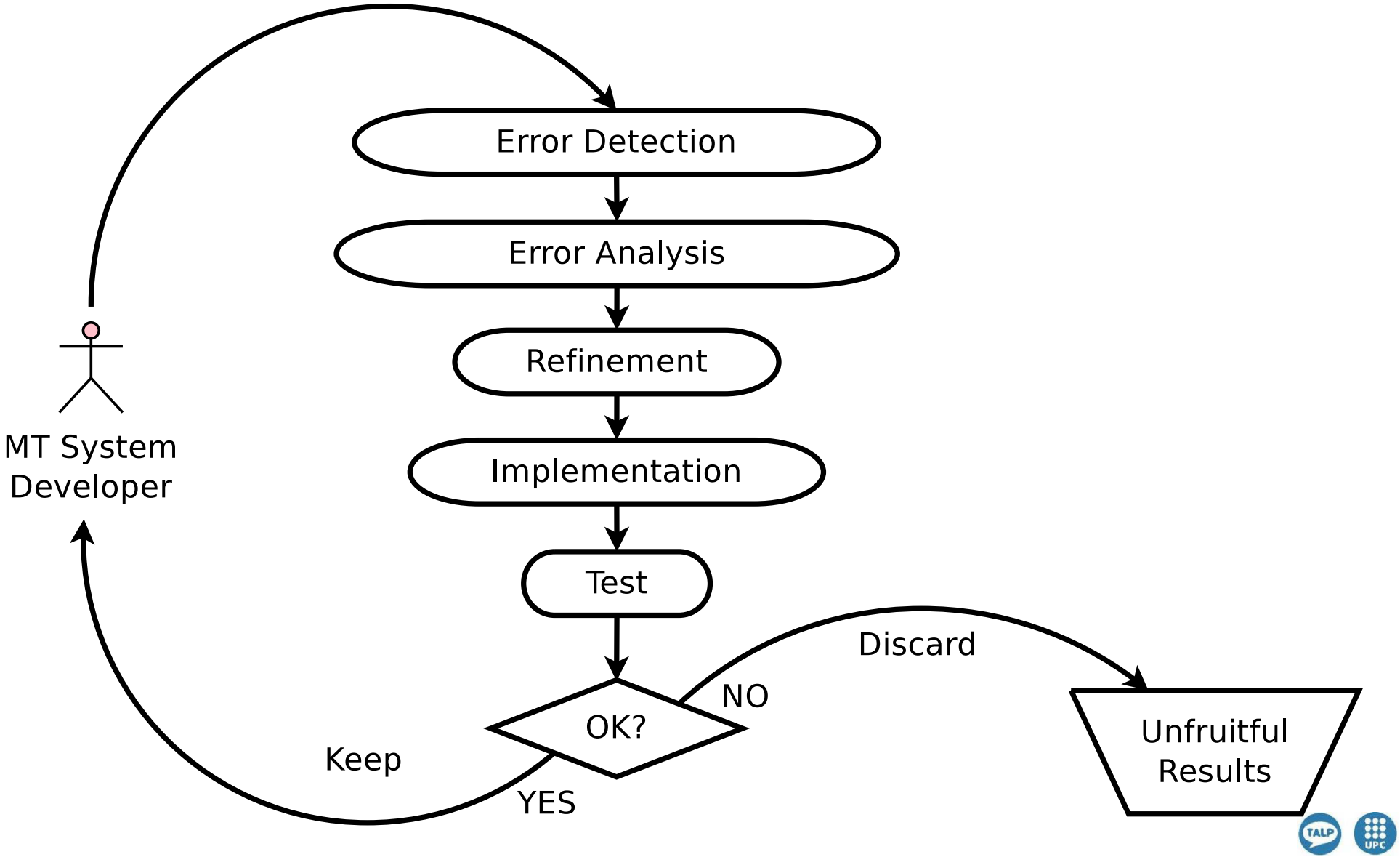
Jesús Giménez

(Advisor, Lluís Màrquez)

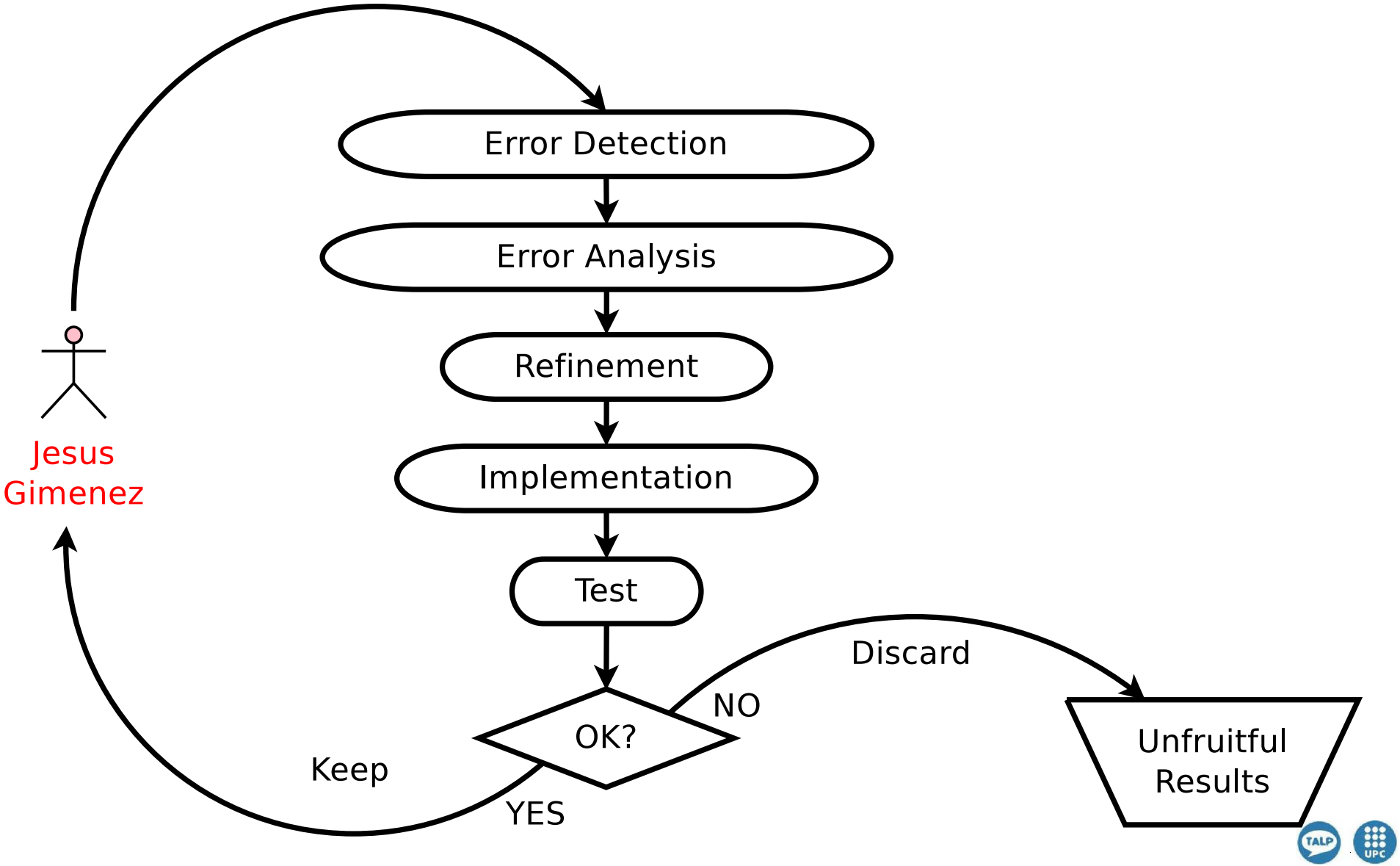
Universitat Politècnica de Catalunya

May 28, 2010

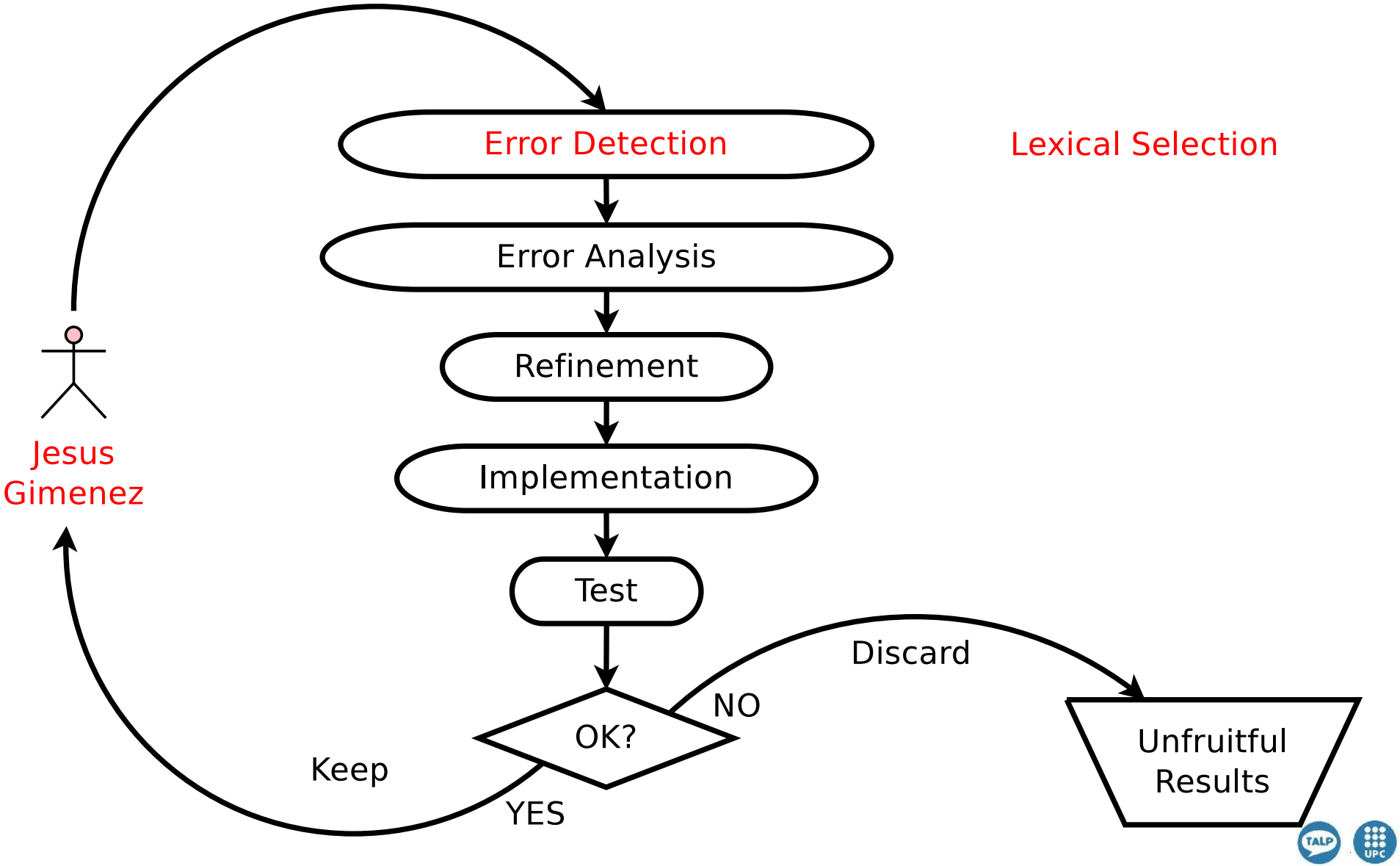
Empirical Machine Translation



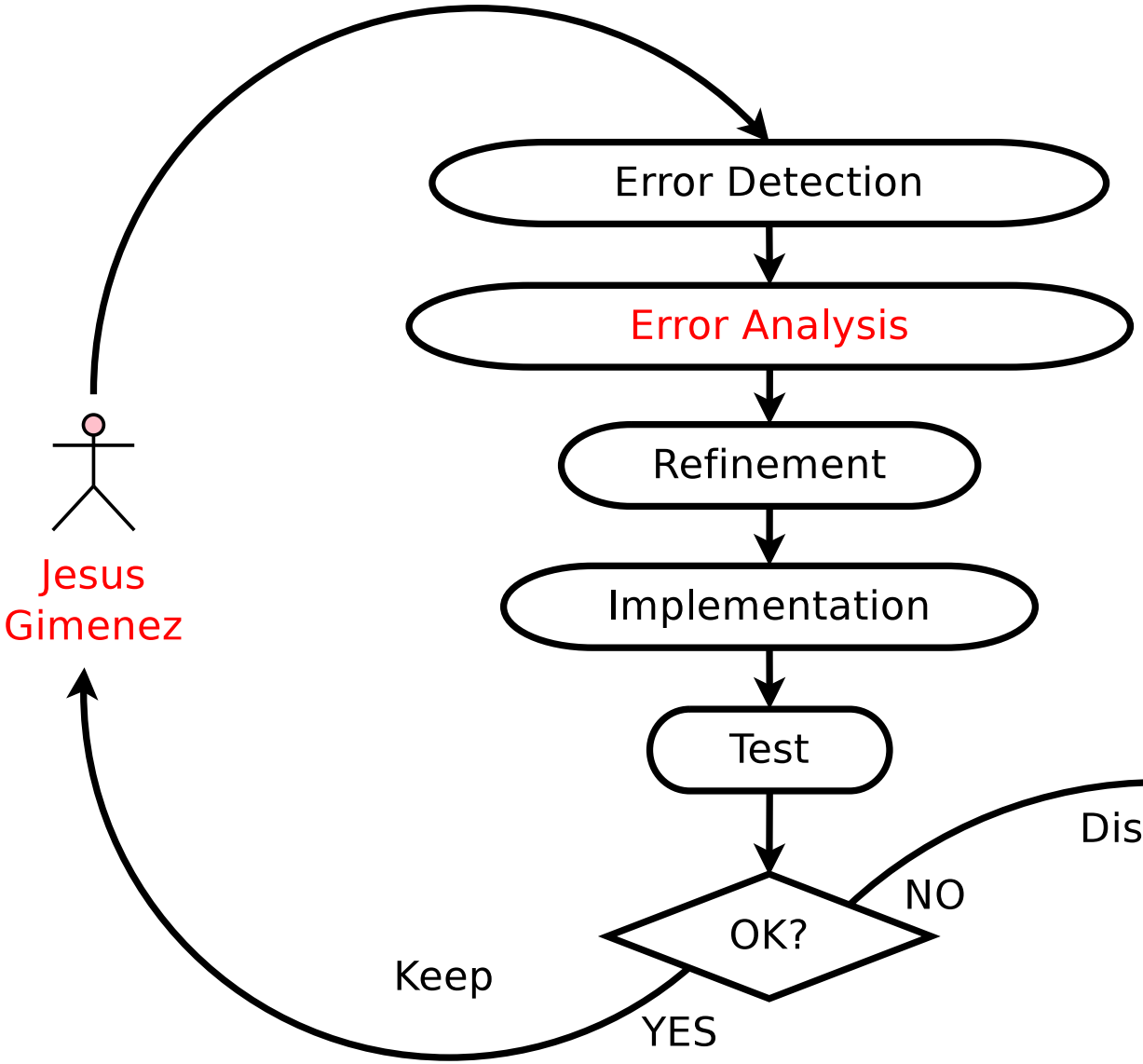
Empirical Machine Translation



Empirical Machine Translation

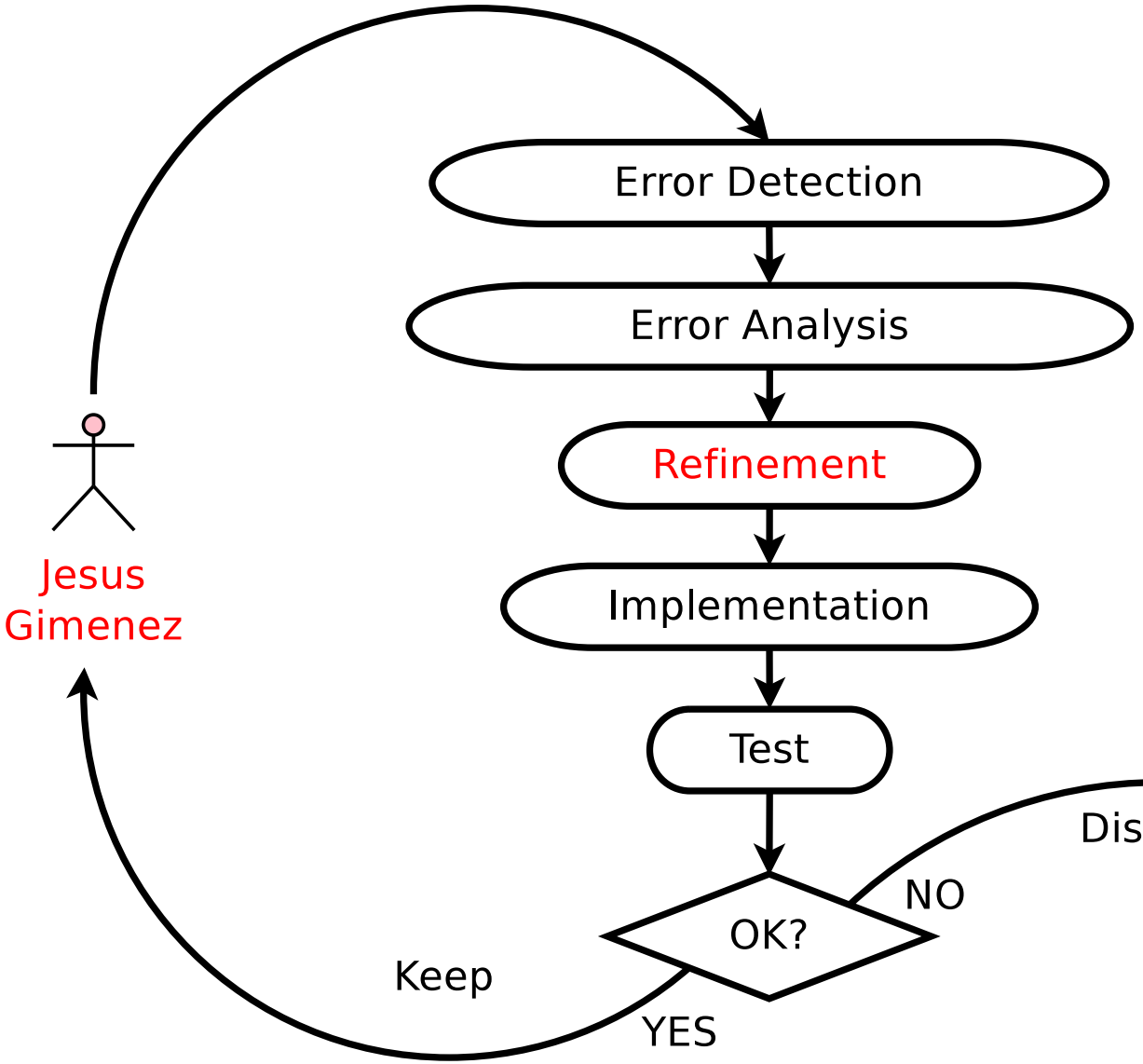


Empirical Machine Translation



No explicit use of linguistic information
+
Source context is not fully exploited

Empirical Machine Translation



Incorporation of
linguistic knowledge
+
Discriminative Learning

Discriminative Phrase Translation

vs. *"A brilliant play by Lionel Messi that produced a wonderful goal"*

→ *"Una brillante jugada de Lionel Messi que resultó en un bello gol"*

Discriminative Phrase Translation

vs. *"A brilliant play by Lionel Messi that produced a wonderful goal"*

→ *"Una brillante jugada de Lionel Messi que resultó en un bello gol"*

Discriminative Phrase Translation

vs. *"A brilliant play by Lionel Messi that produced a wonderful goal"*

$\rightarrow$ *"Una brillante jugada de Lionel Messi que resultó en un bello gol"*

Discriminative Phrase Translation

vs. *"A brilliant play by Lionel Messi that produced a wonderful goal"*

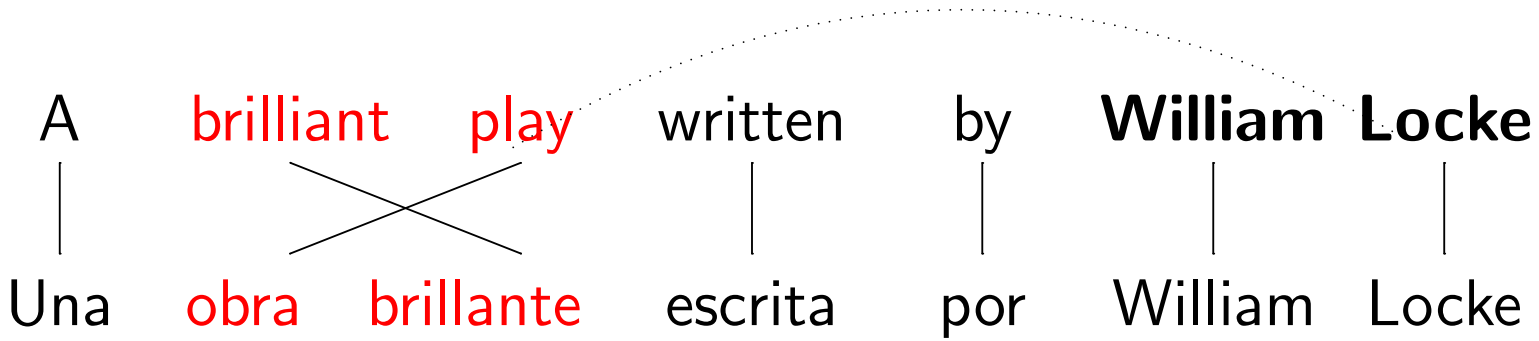
→ *"Una brillante jugada de Lionel Messi que resultó en un bello gol"*

Discriminative Phrase Translation

vs. *"A brilliant play by Lionel Messi that produced a wonderful* **goal***"*

→ *"Una brillante jugada de Lionel Messi que resultó en un bello gol"*

Discriminative Phrase Translation



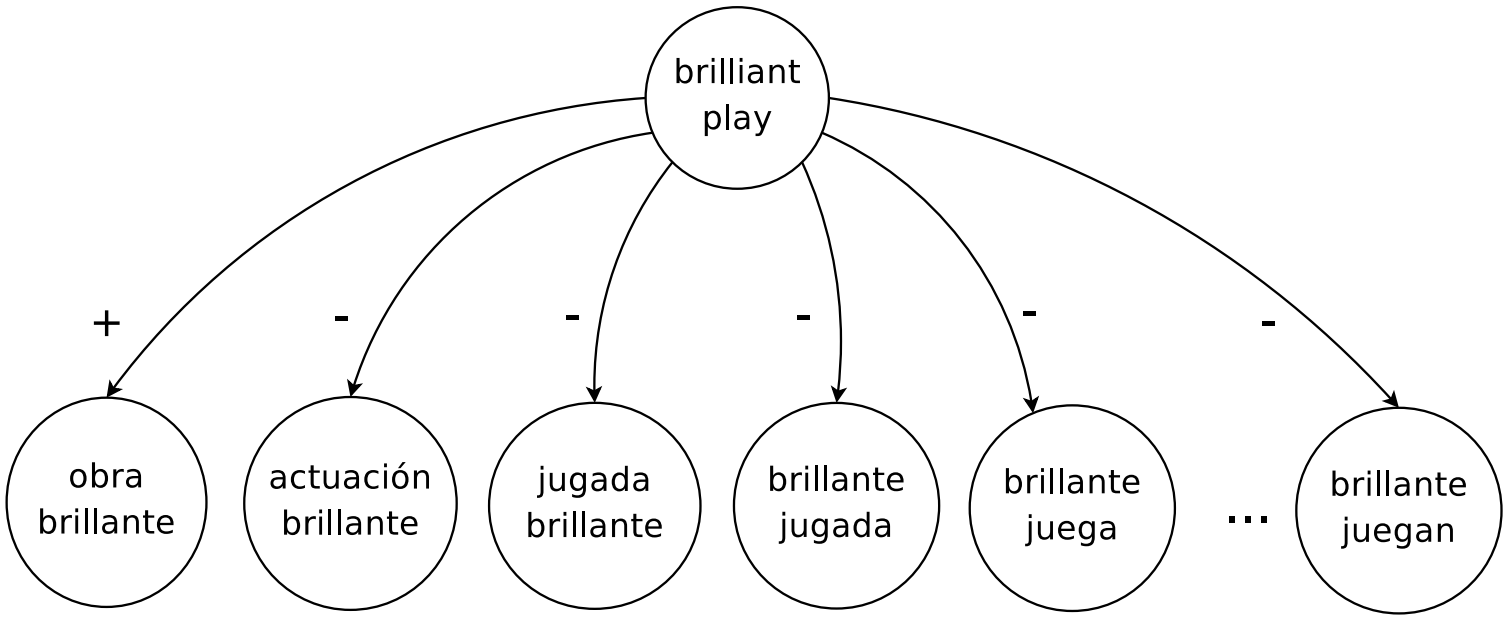
vs. *"A brilliant play by* **Lionel Messi** *that produced a wonderful goal"*

$\rightarrow$ *"Una brillante jugada de Lionel Messi que resultó en un bello gol"*

Discriminative Phrase Translation

Idea

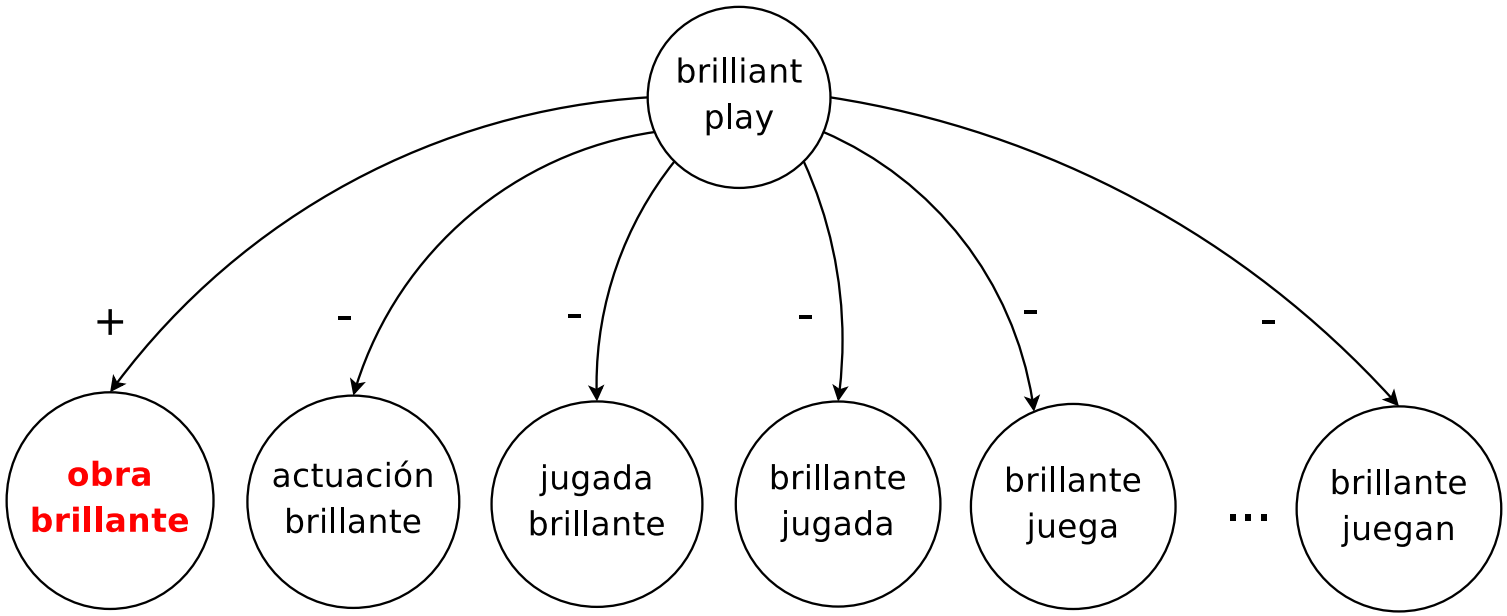
- Use discriminative Machine Learning techniques in SMT to estimate $P(e_i|f_j)$, actually, $P(e_i|f_j, context\_f_j)$
- Translation modeling is addressed as a classification problem



Discriminative Phrase Translation

Idea

- Use discriminative Machine Learning techniques in SMT to estimate $P(e_i|f_j)$, actually, $P(e_i|f_j, context\_f_j)$
- Translation modeling is addressed as a classification problem



Contributions

1. Discriminative Phrase Selection for SMT
   - Shallow-syntactic features
     - word, lemma, parts of speech, chunking
     - local / global context

   → Improved translation quality

2. Domain Dependence in SMT
   - Parliament proceedings → Dictionary definitions
   - Adaptation based on:
     - EuroWordNet
     - out-of-domain data
     - a small amount of in-domain data

   → Improved translation quality

Contributions

1. Discriminative Phrase Selection for SMT
   - Shallow-syntactic features
     - word, lemma, parts of speech, chunking
     - local / global context

   $\longrightarrow$ Improved translation quality
2. Domain Dependence in SMT
   - Parliament proceedings $\longrightarrow$ Dictionary definitions
   - Adaptation based on:
     - EuroWordNet
     - out-of-domain data
     - a small amount of in-domain data

   $\longrightarrow$ Improved translation quality

Contributions

1. Discriminative Phrase Selection for SMT
   - Shallow-syntactic features
     - word, lemma, parts of speech, chunking
     - local / global context

   $\longrightarrow$ Improved translation quality
2. Domain Dependence in SMT
   - Parliament proceedings $\longrightarrow$ Dictionary definitions
   - Adaptation based on:
     - EuroWordNet
     - out-of-domain data
     - a small amount of in-domain data

   $\longrightarrow$ Improved translation quality

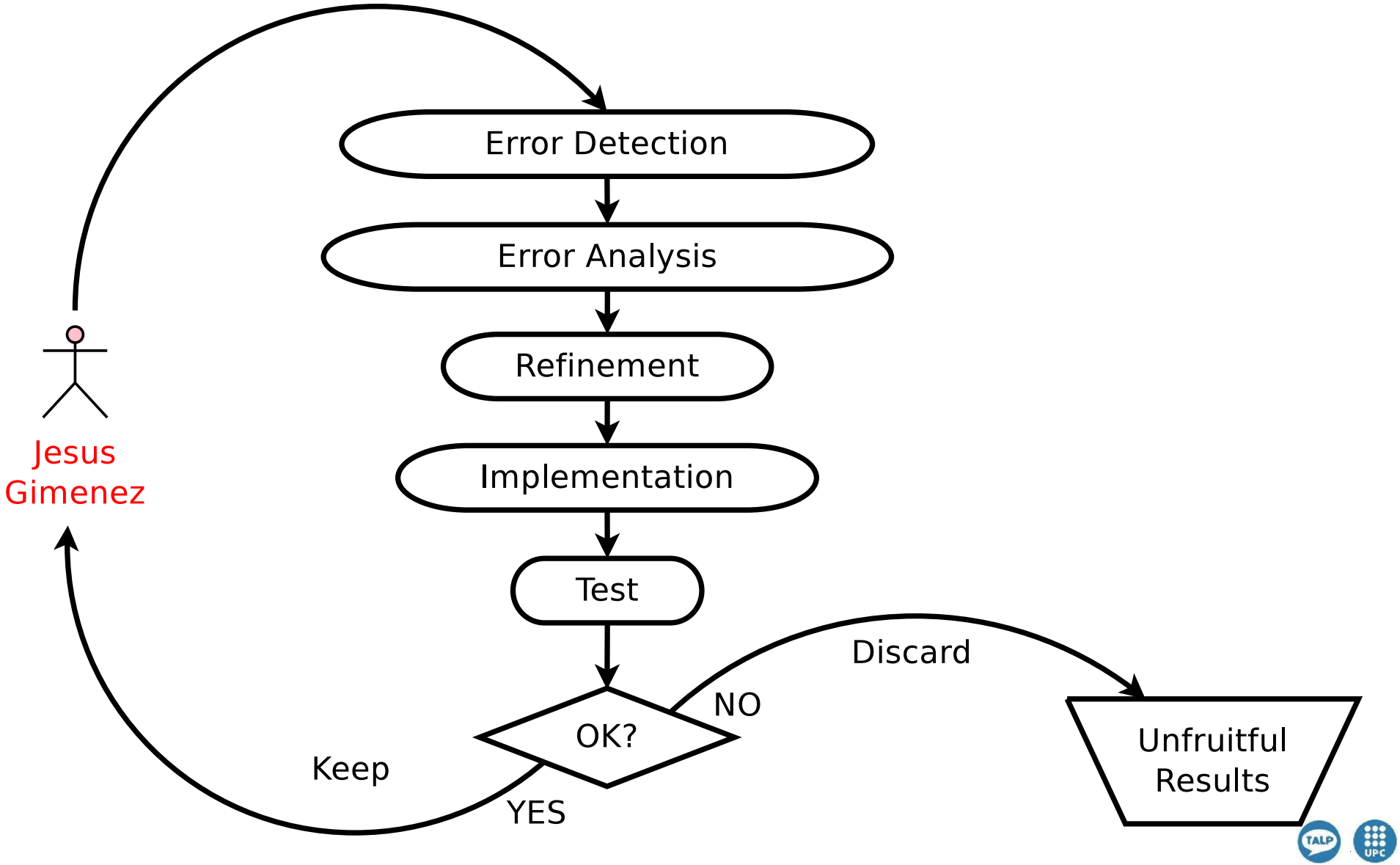
Contributions

1. Discriminative Phrase Selection for SMT
   - Shallow-syntactic features
     - word, lemma, parts of speech, chunking
     - local / global context

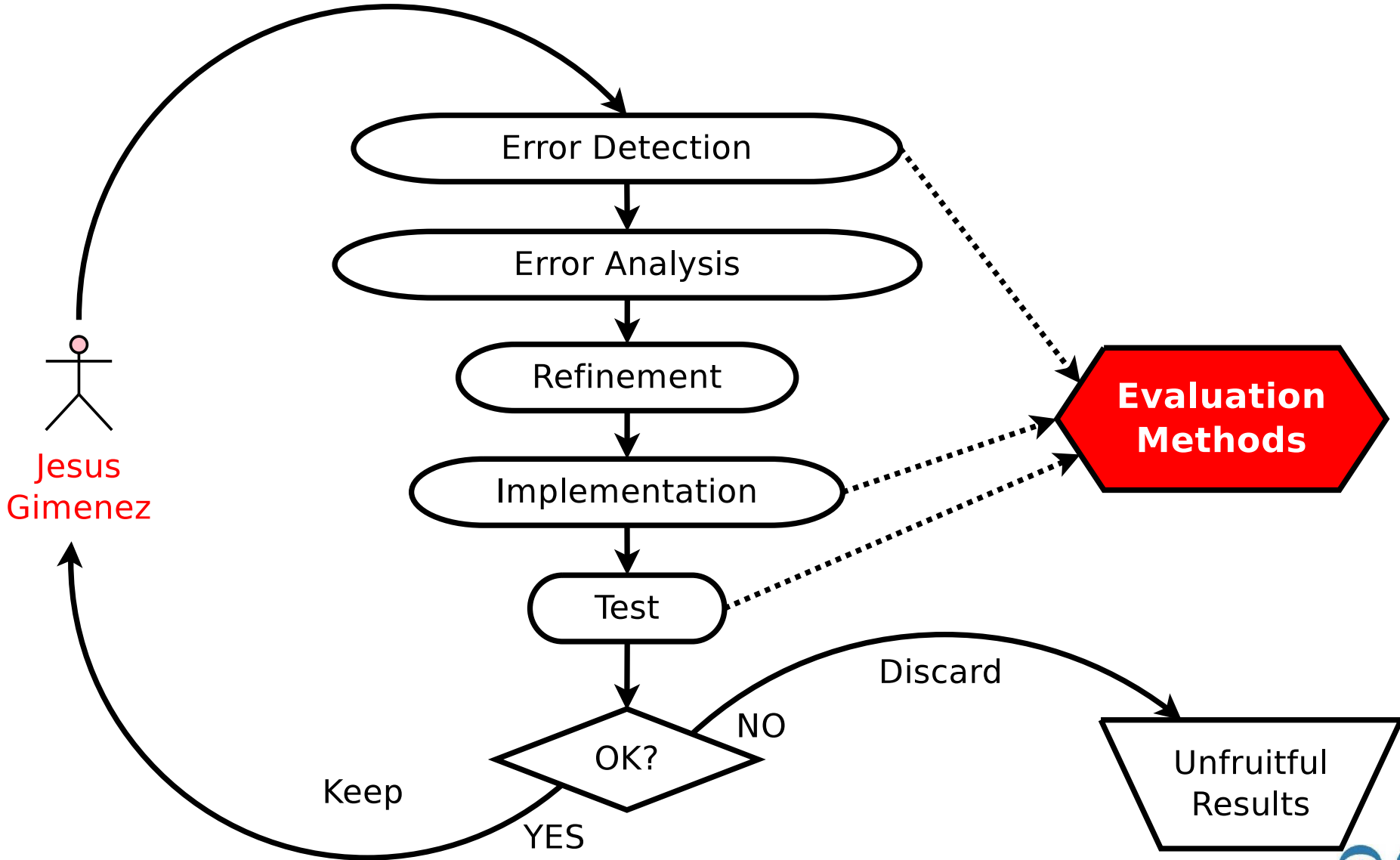
   $\longrightarrow$ Improved translation quality
2. Domain Dependence in SMT
   - Parliament proceedings $\longrightarrow$ Dictionary definitions
   - Adaptation based on:
     - EuroWordNet
     - out-of-domain data
     - a small amount of in-domain data

   $\longrightarrow$ Improved translation quality

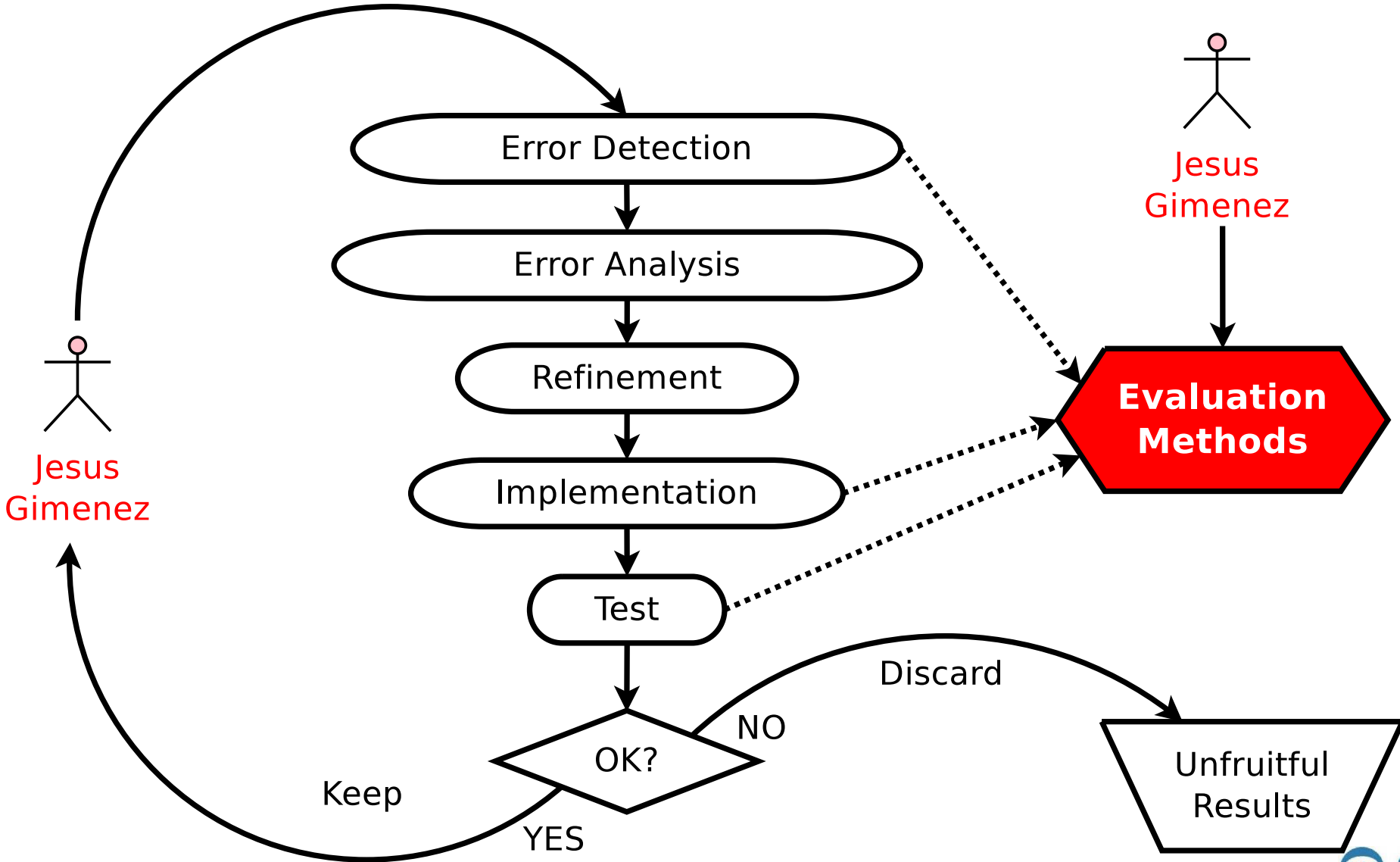
... and its Evaluation



... and its Evaluation



... and its Evaluation



Limits of Lexical Similarity

Candidate Translation	On Tuesday several missiles and mortar shells fell in southern Israel , but there were no casualties .
Reference Translation	Several Qassam rockets and mortar shells fell today, Tuesday , in southern Israel without causing any casualties .

Only one 4-gram in common!

Limits of Lexical Similarity

Candidate Translation	On Tuesday several missiles and mortar shells fell in southern Israel , but there were no casualties .
Reference Translation	Several Qassam rockets and mortar shells fell today, Tuesday , in southern Israel without causing any casualties .

Only one 4-gram in common!

Limits of Lexical Similarity

Candidate Translation	On Tuesday several missiles **and mortar shells fell** in southern Israel , but there were no casualties .
Reference Translation	Several Qassam rockets **and mortar shells fell** today, Tuesday , in southern Israel without causing any casualties .

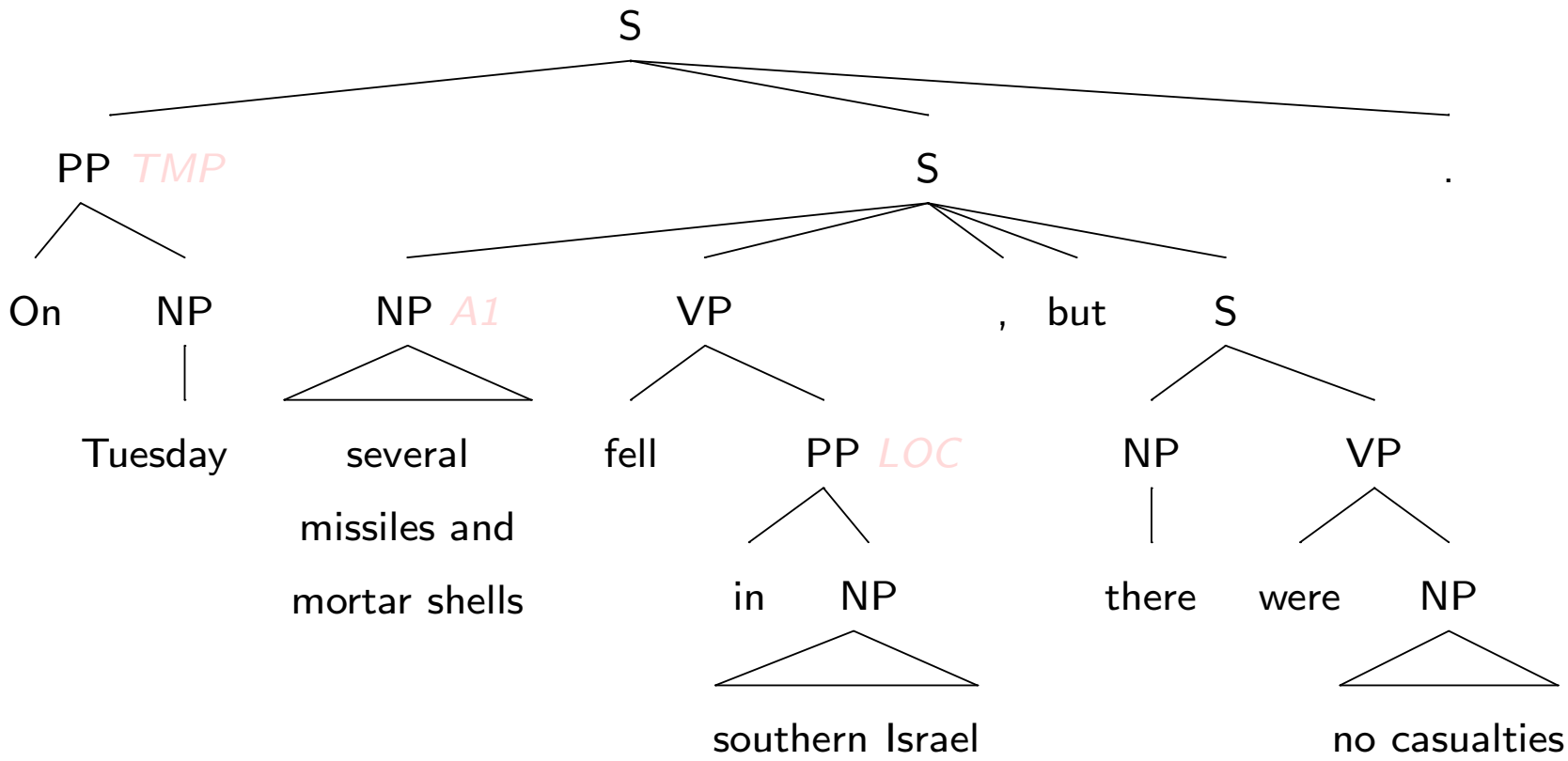
Only one 4-gram in common!

Linguistic Features for Automatic MT Evaluation

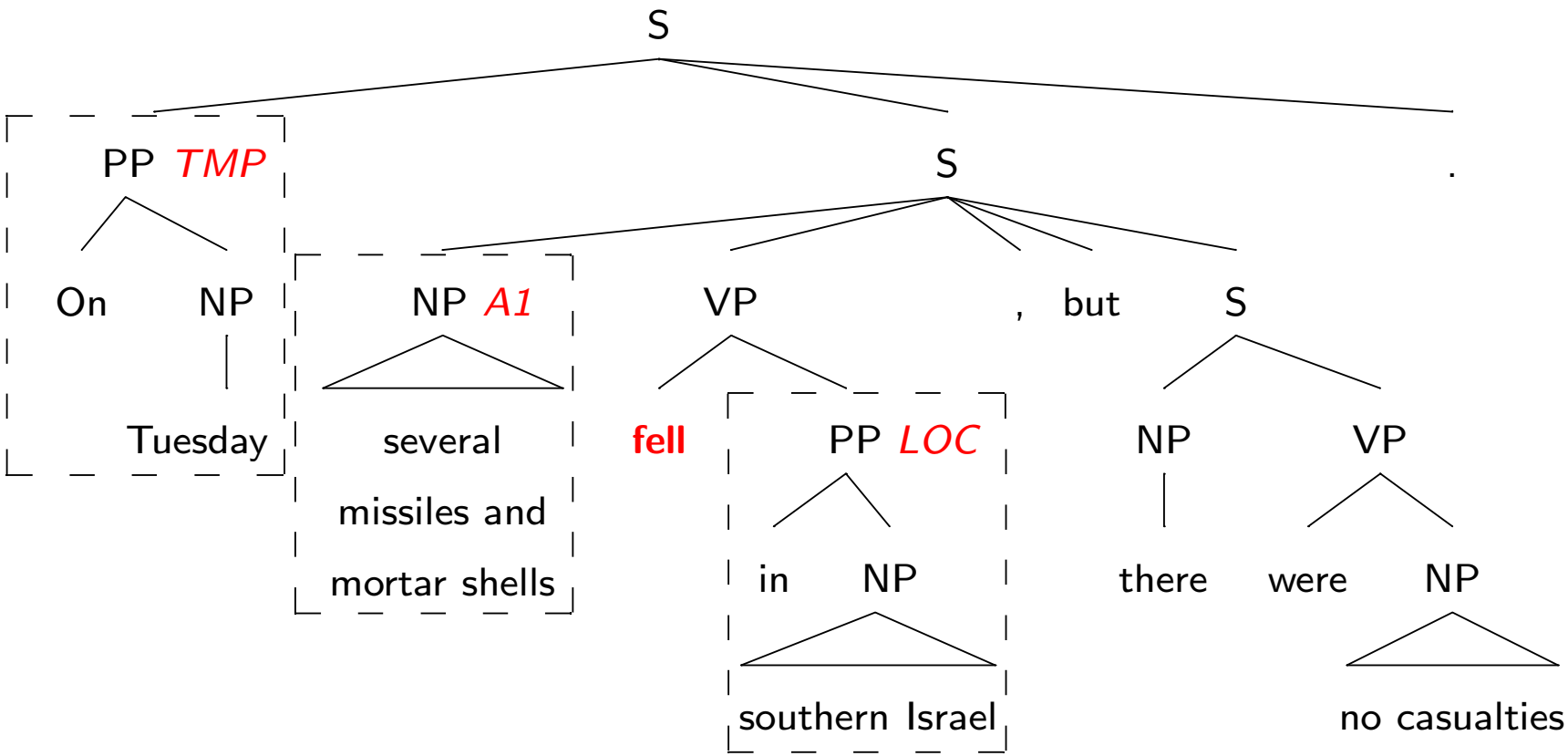
Idea

- Define similarity measures based on deeper linguistic information
  - Compare linguistic structures and their lexical realizations
- Linguistic levels
  - Syntax
    - Parts-of-speech
    - Base phrase chunks
    - Phrase constituents
    - Dependency relationships
  - Semantics
    - Named entities
    - Semantic roles
    - Discourse representations

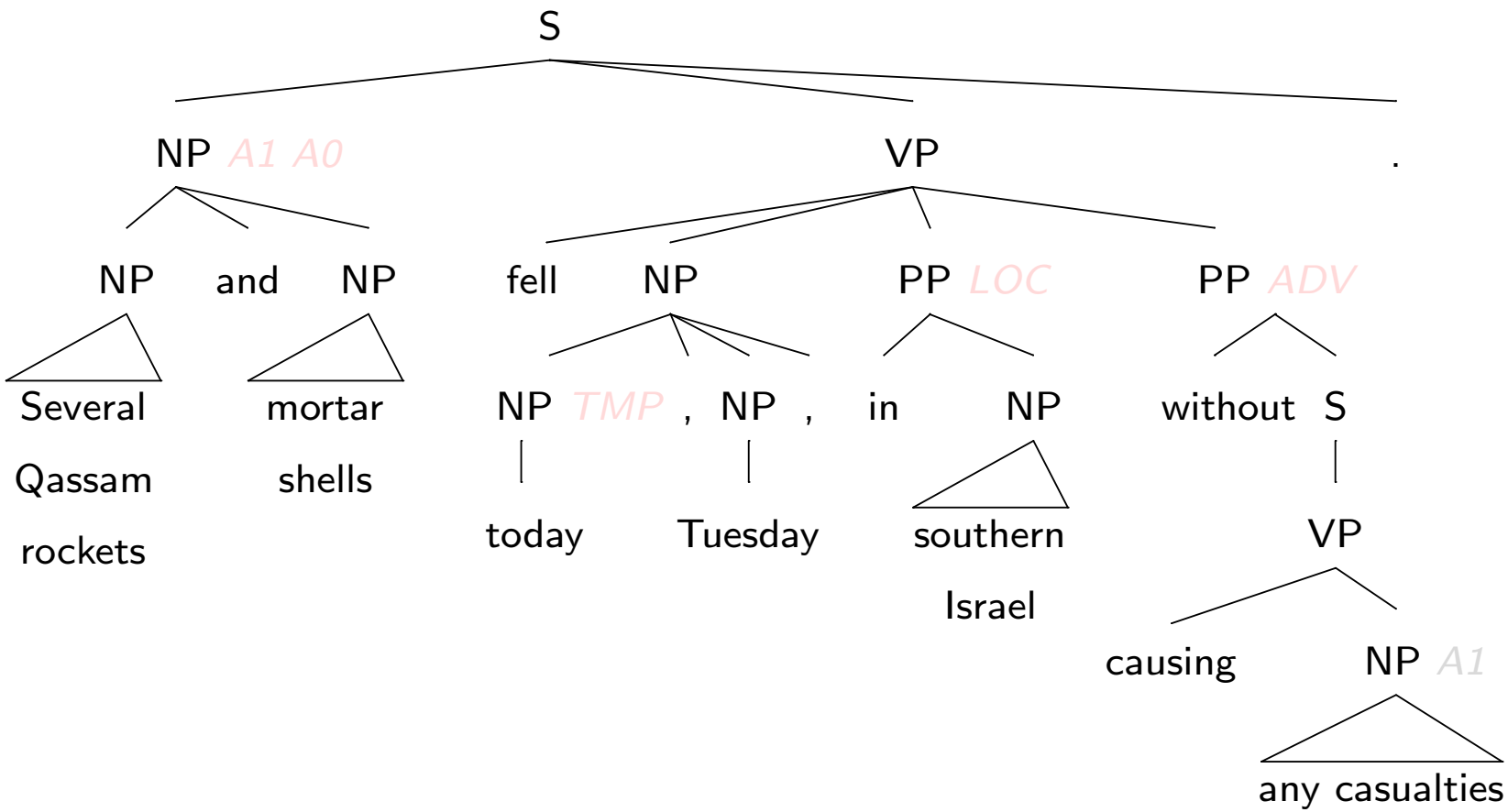
Linguistic Features for Automatic MT Evaluation



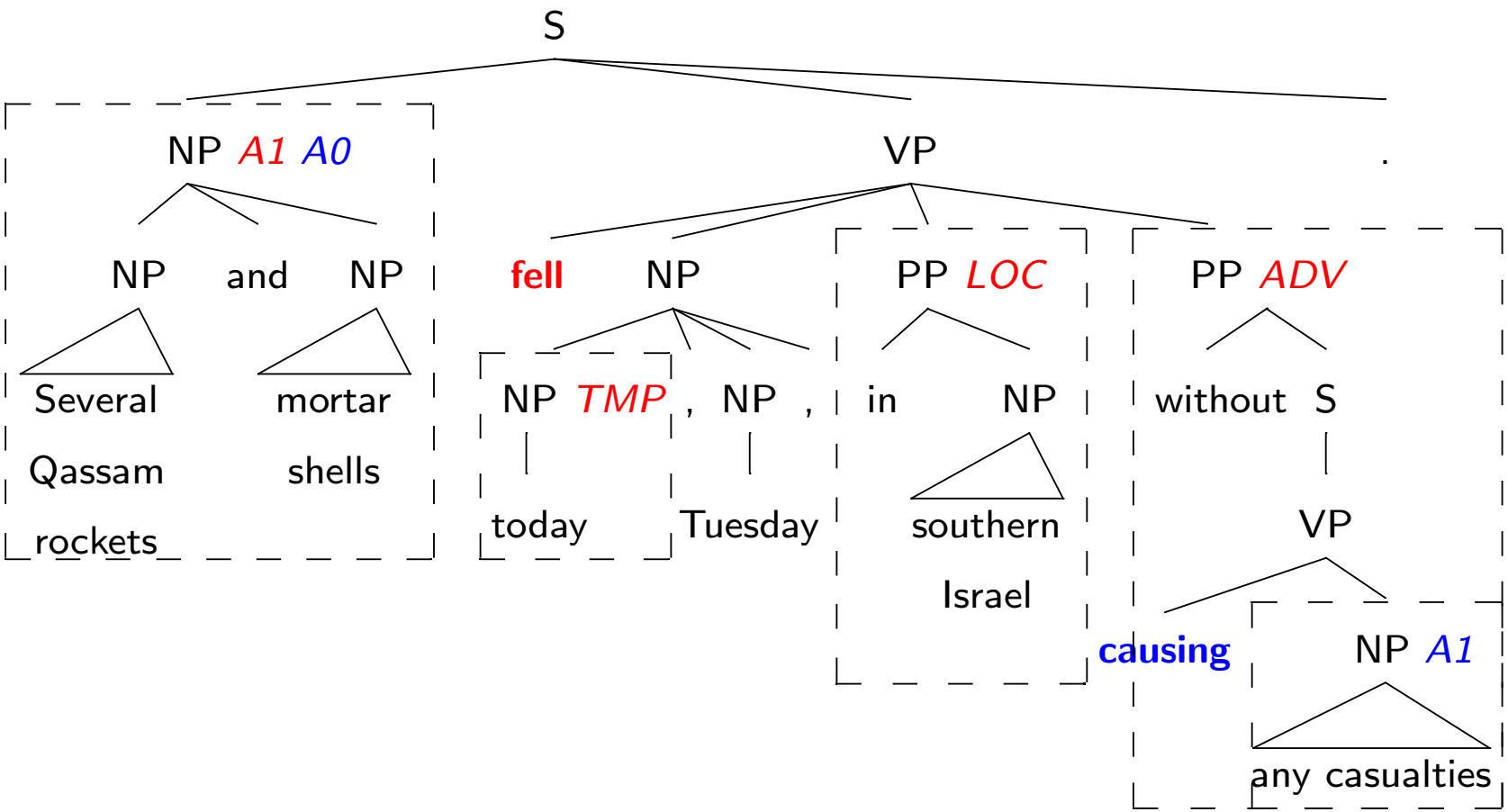
Linguistic Features for Automatic MT Evaluation



Linguistic Features for Automatic MT Evaluation



Linguistic Features for Automatic MT Evaluation



Contributions

1. Linguistic measures provide more reliable rankings when the systems under evaluation are based on different paradigms (statistical vs. rule-based, fully-automatic vs. human-aided)
2. Linguistic measures have proven effective in shared tasks (WMT 2007-2009) both at the system and sentence levels
3. Some linguistic measures suffer a substantial quality decrease at the sentence level (due to parsing errors!)
4. Lexical and Linguistic measures are complementary
   $\longrightarrow$ suitable for being combined!

Contributions

1. Linguistic measures provide more reliable rankings when the systems under evaluation are based on different paradigms (statistical vs. rule-based, fully-automatic vs. human-aided)
2. Linguistic measures have proven effective in shared tasks (WMT 2007-2009) both at the system and sentence levels
3. Some linguistic measures suffer a substantial quality decrease at the sentence level (due to parsing errors!)
4. Lexical and Linguistic measures are complementary
   $\longrightarrow$ suitable for being combined!

Contributions

1. Linguistic measures provide more reliable rankings when the systems under evaluation are based on different paradigms (statistical vs. rule-based, fully-automatic vs. human-aided)
2. Linguistic measures have proven effective in shared tasks (WMT 2007-2009) both at the system and sentence levels
3. Some linguistic measures suffer a substantial quality decrease at the sentence level (due to parsing errors!)
4. Lexical and Linguistic measures are complementary
   $\longrightarrow$ suitable for being combined!

Contributions

1. Linguistic measures provide more reliable rankings when the systems under evaluation are based on different paradigms (statistical vs. rule-based, fully-automatic vs. human-aided)
2. Linguistic measures have proven effective in shared tasks (WMT 2007-2009) both at the system and sentence levels
3. Some linguistic measures suffer a substantial quality decrease at the sentence level (due to parsing errors!)
4. Lexical and Linguistic measures are complementary
   $\longrightarrow$ suitable for being combined!

Contributions

1. Linguistic measures provide more reliable rankings when the systems under evaluation are based on different paradigms (statistical vs. rule-based, fully-automatic vs. human-aided)
2. Linguistic measures have proven effective in shared tasks (WMT 2007-2009) both at the system and sentence levels
3. Some linguistic measures suffer a substantial quality decrease at the sentence level (due to parsing errors!)
4. Lexical and Linguistic measures are complementary
   $\longrightarrow$ **suitable for being combined!**

Acknowledgements

1. Spanish Government
   - Ministry of Science and Technology
   - Ministry of Education
2. Organizers and participants of the NIST, WMT and IWSLT Evaluation Campaigns
3. A number of NLP researchers worldwide sharing their software
4. Enrique Amigó and German Rigau

Acknowledgements

1. Spanish Government
   - Ministry of Science and Technology
   - Ministry of Education
2. Organizers and participants of the NIST, WMT and IWSLT Evaluation Campaigns
3. A number of NLP researchers worldwide sharing their software
4. Enrique Amigó and German Rigau

Acknowledgements

1. Spanish Government
   - Ministry of Science and Technology
   - Ministry of Education
2. Organizers and participants of the NIST, WMT and IWSLT Evaluation Campaigns
3. A number of NLP researchers worldwide sharing their software
4. Enrique Amigó and German Rigau

Acknowledgements

1. Spanish Government
   - Ministry of Science and Technology
   - Ministry of Education
2. Organizers and participants of the NIST, WMT and IWSLT Evaluation Campaigns
3. A number of NLP researchers worldwide sharing their software
4. Enrique Amigó and German Rigau

Empirical Machine Translation and its Evaluation

EAMT Best Thesis Award 2008

Thanks!