Integration of SYSTRAN MT systems in an open workflow

Mats Attnäs¹  Pierre Senellart¹,²  Jean Senellart¹

¹ SYSTRAN® ² INRIA

MT Summit X
September 15th, 2005
Introduction

New architecture

Applications

Conclusion

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- \( \approx 35,000,000 \) on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
## Facts

- **35-year-old company**
- **40+ language pairs**
- **20+ different languages**
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- **≈ 35,000,000** on-line translations a day
- **Large** range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
# SYSTRAN

## Facts
- **35-year-old company**
- **40+ language pairs**
- **20+ different languages**
- **Translation service for major portals** (Babelfish, Google, Yahoo!...)
  - \( \approx 35,000,000 \) on-line translations a day
- **Large range of products** (from PDA to Web servers)
- **Services for corporate and institutions** (Ford, Cisco, US Gov, EU...)
SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- \( \approx 35,000,000 \) on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- $\approx 35,000,000$ on-line translations a day
- **Large** range of products (from PDA to Web servers)
- Services for **corporate and institutions** (Ford, Cisco, US Gov, EU...)
SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- \( \approx 35,000,000 \) on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
Recent activities

- **New Generation (NG) systems**: Arabic → English, Swedish → English
- **Tools for translators**: (translation memory support, bilingual terminology extraction...)
- **New Language Pairs**: (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- **SYSTRAN Lite on PDA**
Recent activities

- **New Generation (NG) systems**: Arabic $\rightarrow$ English, Swedish $\rightarrow$ English
- **Tools for translators**: (translation memory support, bilingual terminology extraction...) 
- **New Language Pairs**: (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- **SYSTRAN Lite on PDA**
Recent activities

- **New Generation (NG) systems**: Arabic $\rightarrow$ English, Swedish $\rightarrow$ English
- Tools for **translators** (translation memory support, bilingual terminology extraction...)
- **New Language Pairs**: (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- SYSTRAN Lite on PDA
Recent activities

- **New Generation** (NG) systems: Arabic → English, Swedish → English
- Tools for **translators** (translation memory support, bilingual terminology extraction…)
- **New** Language Pairs (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- **SYSTRAN Lite** on PDA
MT System Complexity

- \((nb_{rules}, size_{dictionary}) \times nb_{LPS}\)

- High flexibility required

- Stability

- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

\[ \implies \text{unavoidable saturation point?} \]
MT System Complexity

- \((nb_{rules}, size_{dictionary}) \times nb_{LPS}\)

- **High flexibility** required

- Stability

- **Intrinsic complexity** of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

\[\implies \text{unavoidable saturation point?}\]
MT System Complexity

- \((nb_{rules}, size_{dictionary}) \times nb_{LPs}\)
- High flexibility required
- Stability
  - Intrinsic complexity of language description
    - Rules and exceptions
    - Rule interaction
    - Endless dictionary completion
  \[\rightarrow\] unavoidable saturation point?
MT System Complexity

- \((nb_{rules}, size_{dictionary}) \times nb_{LPs}\)
- High flexibility required
- Stability
- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion
  \[\Rightarrow\] unavoidable saturation point?

Atnäs, Senellart, Senellart

Integration of SYSTRAN MT systems in an open workflow
MT System Complexity

\[(\text{nb}_{\text{rules}}, \text{size}_{\text{dictionary}}) \times \text{nb}_{\text{LPS}}\]

- High flexibility required
- Stability
- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

\[\Rightarrow\] unavoidable saturation point?
A black box system is not enough!

Need of interaction

<table>
<thead>
<tr>
<th>Level</th>
<th>Issue</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguist</td>
<td>Why do I get this?</td>
<td>traceability, entry point</td>
</tr>
<tr>
<td>Translator</td>
<td>Where should I look?</td>
<td>risk area, alternatives, resource coverage</td>
</tr>
<tr>
<td>Author</td>
<td>How will it translate?</td>
<td>MT-in-mind authoring</td>
</tr>
<tr>
<td>End user</td>
<td>What can I change?</td>
<td>resource interaction</td>
</tr>
</tbody>
</table>

Evolution requires “opening”.

Atnäs, Senellart, Senellart

Integration of SYSTRAN MT systems in an open workflow
A black box system is not enough!

**Need of interaction**

<table>
<thead>
<tr>
<th>Level</th>
<th>Issue</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguist</td>
<td>Why do I get this?</td>
<td>traceability, entry point</td>
</tr>
<tr>
<td>Translator</td>
<td>Where should I look?</td>
<td>risk area, alternatives, resource coverage</td>
</tr>
<tr>
<td>Author</td>
<td>How will it translate?</td>
<td>MT-in-mind authoring</td>
</tr>
<tr>
<td>End user</td>
<td>What can I change?</td>
<td>resource interaction</td>
</tr>
</tbody>
</table>

E**VOLUTION REQUIRES “OPENING”**.
A black box system is not enough!

### Need of interaction

<table>
<thead>
<tr>
<th>Level</th>
<th>Issue</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguist</td>
<td>Why do I get this?</td>
<td>traceability, entry point</td>
</tr>
<tr>
<td>Translator</td>
<td>Where should I look?</td>
<td>risk area, alternatives, resource coverage</td>
</tr>
<tr>
<td>Author</td>
<td>How will it translate?</td>
<td>MT-in-mind authoring</td>
</tr>
<tr>
<td>End user</td>
<td>What can I change?</td>
<td>resource interaction</td>
</tr>
</tbody>
</table>

**Evolution requires “opening”**.
Outline

1. Introduction
2. New architecture
   - Code modernization
   - Process Modularization
   - Data structures
   - Natural Language Oriented Matching
3. Applications
   - Input Simplification
   - Control mechanisms
   - New Language Pairs
4. Conclusion
From specific transliterations and encodings to standard Unicode

From low-level paradigms to high-level object-oriented code

⇒

- Redesign, Modularization
- High-level linguistic data structures

Joining parallel effort of New Generation engines (high-level, declarative)
Code modernization

- From specific transliterations and encodings to standard **Unicode**
- From low-level paradigms to **high-level object-oriented code**
  - Redesign, Modularization
  - High-level linguistic data structures
- Joining parallel effort of **New Generation engines** (high-level, declarative)

Integration of SYSTRAN MT systems in an open workflow
Code modernization

- From specific transliterations and encodings to standard Unicode
- From low-level paradigms to high-level object-oriented code

  ⇒
  - Redesign, Modularization
  - High-level linguistic data structures

- Joining parallel effort of New Generation engines (high-level, declarative)
Code modernization

- From specific transliterations and encodings to standard Unicode
- From low-level paradigms to high-level object-oriented code
  \[\Rightarrow\]
  - Redesign, Modularization
  - High-level linguistic data structures
- Joining parallel effort of New Generation engines (high-level, declarative)
A flow of XML data to convey more information:
- source text and history
- intermediate status
- document structure information
- internal/external markup
- confidence information

Processed by independent, interoperable agents:
- Linguistics (Analysis/Transfer/Generation)
- Generic (Preprocessing/Postprocessing)
XML Workflow & Modules

- A flow of XML data to convey **more information**
  - source text and history
  - intermediate status
  - document structure information
  - internal/external markup
  - confidence information

- Processed by **independent, interoperable** agents
  - **Linguistics** (Analysis/Transfer/Generation)
  - **Generic** (Preprocessing/Postprocessing)
Workflow Schema

Integration of SYSTRAN MT systems in an open workflow
Atnäs, Senellart, Senellart

Integration of SYSTRAN MT systems in an open workflow
Integration of SYSTRAN MT systems in an open workflow
From relation graphs to syntagm trees

The director comments on the making of the film.
The director comments on the making of the film.
The director comments on the making of the film.
Declarative linguistic rules

\[
\text{NP} \xrightarrow{\text{CONJ:+coordinate}} \text{NP} \xrightarrow{\text{NP:+plural}}
\]

Atnäs, Senellart, Senellart

Integration of SYSTRAN MT systems in an open workflow
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

All of this requires "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas...)
- On-the-fly spellcheck, normalization...
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

All of this requires "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas...)
- On-the-fly spellcheck, normalization...
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES “SMART MATCHING”.

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas...)
- On-the-fly spellcheck, normalization...
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

All of this requires “SMART MATCHING”.

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas…)
- On-the-fly spellcheck, normalization…
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

**All of this requires “SMART MATCHING”.**

- XML structured matching
  - Low-level lookup operators (numbers, dates, chemical formulas . . .)
  - On-the-fly spellcheck, normalization . . .
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

All of this requires “SMART MATCHING”.

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas...)
- On-the-fly spellcheck, normalization...
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

All of this requires “SMART MATCHING”.

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas...)
- On-the-fly spellcheck, normalization...
Spellcheck
Outline

1. Introduction
2. New architecture
   - Code modernization
   - Process Modularization
   - Data structures
   - Natural Language Oriented Matching
3. Applications
   - Input Simplification
   - Control mechanisms
   - New Language Pairs
4. Conclusion
Linguistic Improvement by Input Simplification

Let specialized agents bring their expertise.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making
LET SPECIALIZED AGENTS BRING THEIR EXPERTISE.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making
Linguistic Improvement by Input Simplification

Let specialized agents bring their expertise.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making
Linguistic Improvement by Input Simplification

Let specialized agents bring their expertise.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making
LET SPECIALIZED AGENTS BRING THEIR EXPERTISE.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making
Interacting with the user

SYSTRAN Translation Project Manager (STPM)

The director comments on the making of the film.

Le directeur présente ses observations sur la fabrication du film.

Atnäs, Senellart, Senellart

Integration of SYSTRAN MT systems in an open workflow
With more than 1.5 million installed, the Cisco 2500 series is one of the most popular solutions for a wide range of cost-effective configurations, including dual LAN, integrated router/hub, and integrated access server models.

For example, integrated call switching and call handling features enable small or branch offices to use their Cisco access solution for call handling and remote access instead of having to invest in a PBX system.

Avec plus de 1.5 million étant installé, les 2500 séries de Cisco sont l’une des solutions les plus populaires pour un éventail de configurations, y compris le LAN duel, de routeur/hub intégrés, et de modèles intégrés de serveur d’accès.

Par exemple, la commutation d’appel et les dispositifs intégrés de manipulation d’appel permettent succursales de petites ou d’employer leur solution d’accès de Cisco pour la manipulation d’appel et l’accès à distance au lieu de devoir investir dans un système de PBX.
Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?
Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?
Developing new language pairs

Analysis, Transfer, Generation:
- Independence
- Genericity
- Maturity

Challenge
40 new cross LPs in one year?
Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?
Developing new language pairs

Analysis, Transfer, Generation:
- Independence
- Genericity
- Maturity

Challenge
40 new cross LPs in one year?
Summary

- Still a complex system
- But clear entry points
- Open to external approaches
- Strengthened interactivity with linguist/translator/author/user
Summary

- Still a **complex system**
- But clear **entry points**
- Open to **external approaches**
- Strengthened **interactivity** with linguist/translator/author/user
Summary

- Still a complex system
- But clear entry points
- Open to external approaches
- Strengthened interactivity with linguist/translator/author/user
Summary

- Still a **complex system**
- But clear **entry points**
- Open to **external** approaches
- Strengthened **interactivity** with linguist/translator/author/user
Perspectives

- Effective use of alternative approaches in v6
  - To know the system’s hesitations and help it decide (rules weighting)
  - Smaller agents
  - Improved quality, New LPs!
Perspectives

- Effective use of alternative approaches in v6
- To know the system’s hesitations and help it decide (rules weighting)
- Smaller agents
- Improved quality, New LPs!
Effective use of alternative approaches in v6

To know the system’s hesitations and help it decide (rules weighting)

Smaller agents

Improved quality, New LPs!
Effective use of alternative approaches in v6

To know the system’s hesitations and help it decide (rules weighting)

Smaller agents

Improved quality, New LPs!