Statistical Machine Translation

Marcello Federico
FBK-irst Trento, Italy
Galileo Galilei PhD School - University of Pisa

Pisa, 7-19 May 2008

Part I: Overview of Machine Translation

- Introduction to MT
- Approaches
- Statistical MT Framework
- Phrase-based SMT
- MT Evaluation
- State-of-the-art
- Examples

Machine Translation

A general definition:

> **Machine translation**, sometimes referred to by the acronym MT, is a sub-field of computational linguistics that **investigates the use of computer software to translate text or speech from one natural language to another**. (Wikipedia, the free encyclopedia.)

My Definition

MT investigates the translation of "standard" language that can be systematically observed in ordinary communication – e.g. conversations, news, speeches, business letters, user manuals, etc. –. MT as a discipline is not interested in the translation of literature genres that express creative and sophisticated use of language. For several reasons, such kind of language is simply out of the scope of MT.

For a very interesting introduction to issues related to the translation of literature works see U. Eco, "Dire quasi la stessa cosa. Esperienze di traduzione.", Bompiani 2003.

Introduction to MT

Why is Machine Translation so Difficult?

High quality human translation implies:

- deep and rich understanding of source language and text
- sophisticated and creative command of target language

Nowadays, feasible goals for machine translation are only tasks:

- for which a rough translation is adequate (gist translation)
- where a human post-editor can improve MT output (CAT)
- focusing on small linguistic domains (translators on PDAs)

In general, difficulty of translating depends on how similar the target and source languages are in their vocabulary, grammar, and conceptual structure.

Differences and Similarities of Languages

- **Universal communicative role** of language
  – names for people, words for talking about women, men, children
  – every language seems to have nouns and verbs
- **Differences/similarities across large classes of languages**:
  – Morphological: one vs. many morphemes per words, agglutination vs. fusion
  – Syntactical: Subj-Verb-Obj structure (E) vs. SOV (J) vs. VSO (Irish)
  – Semantical: direction/manner of motion indicated by verb/satellites
  `the bottle` <u>`floated`</u> <u>`out`</u> (E) → `la botella` <u>`salió`</u> <u>`flotando`</u> (S)
- **Differences in specificity**, often peculiar to single languages:
  – Lexical: `informatique` (F) → `computer science` (E)
  – Syntactical: `she` <u>`likes`</u> `to sing` (E,v) → `sie singt` <u>`gerne`</u> (D,adv)
  – Semantical: `wall` (E) → `Wand/Mauer` (G, inside/outside)
- **Cultural Differences**: philosophical argument=is translation possible at all?

Differences in Specificity

English	brother	Japanese	otooto (younger)
		Japanese	aniisan (older)
English	is	Japanese	isu (subject is animate)
		Japanese	aru (subject is not animate)
English	know	French	connaître (be acquainted with)
		French	savoir (know a proposition)
English	they	French	ils (masculine)
		French	elles (feminine)
German	Berg	English	hill
		English	mountain

A Brief History of Machine Translation

before 1900	various suggestions about "mechanic" translation
1933	French Patent by George Artsouni:
	storage device on paper tape to find translations of words
	Russian Patent by **Petr Petrovich Troyanskii**:
	lexical-syntactic transfer (base-forms+syntactic functions)
1949	memorandum by **Warren Weaver** (and Andrew D. Booth):
	cryptography methods, statistical methods, Shannon's theory
1951	First research position on MT at MIT
1954	rule-based MT project by Georgetown U. + IBM:
	public demo Russian to English (Vocab: 250 words, Grammar: 6 rules)
1955	U. Leningrad: interlingua as artificial language

A Brief History of Machine Translation

1956-1966	**large scale funding in US**: high expectation & disillusion
1957	**Peter Toma** starts building Systran
1958	U. Washington, IBM : word-for-word approach
	Russian-English system for US Air Force (up to 1970)
1960	RAND corp. rough translation with statistical approach
1961	U. Georgetown (+ P. Toma) Russian to English demo
	rule based (more levels of analysis)
around 1960	MIT and U. Texas work on syntactic transfer approach
1967	**ALPAC report**: US funding drastically reduced for 10 years
1970-1981	U. Montreal, TAUM project: rule-based, logic-programming
	success with weather forecasts, failure with aviation manuals
1960-1971	U. Texas and U. Grenoble work on interlingua approach, logic
1975	interlingua looses interest

A Brief History of Machine Translation

1980 -	Rule based transfer and new interlingua approaches based on linguistic theories, logic programming, AI
1990 -	**Rule based MT dominance is broken** Statistical alignment models for French-English (IBM) Example-based translation (Sato and Nagao, Japan)
1990 -	Speech translation projects: limited domains ATR, Kyoto: automatic telephony research CSTAR consortium (US, Europe, Asia) Verbmobil project (Germany)
2000 -	**Unrestricted Spoken Language Translation** Automatic evaluation metrics for MT (IBM) TIDES (US): written news Chi/Ara to Eng GALE (US): broadcast news Chi/Ara to Eng TC-STAR (EU): news Chi to Eng speeches Spa-Eng

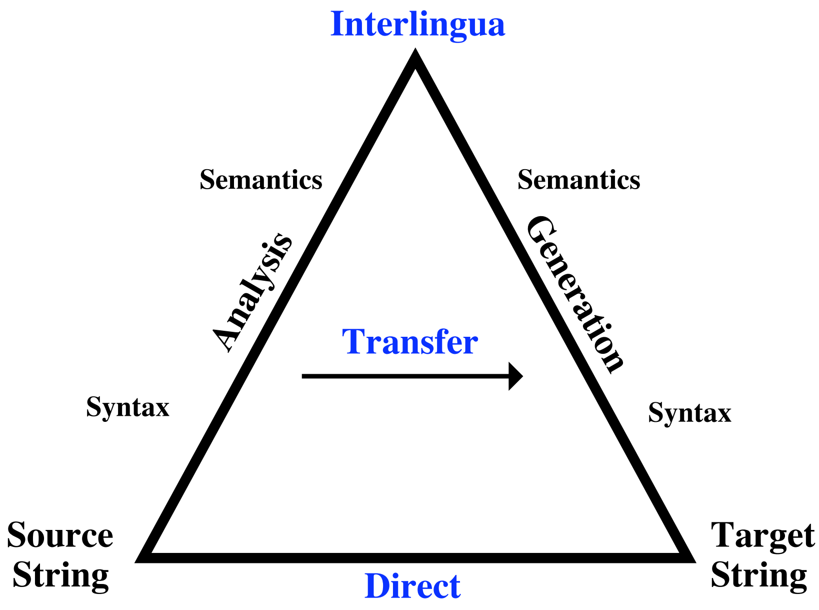
Approaches to MT

Rough classification according to **employed linguistic representations**:

- **Direct model**: translate and re-order single words or n-grams
  – basically, no linguistic representation is used
- **Transfer model**: use explicit knowledge about language differences
  – analyze lexical and syntactic structure of source sentence
  – transfer structures from source to target language
  – generate corresponding sentence in the target language
- **Interlingua model**: extract the meaning and express it in the target language
  – analyze lexical, syntactical and semantical structure of source sentence
  – interpret the meaning into a canonical interlingua
  – generate the target sentence from the interlingua

Notice: required knowledge for the interlingua approach grows linearly with number of languages, rather than to the square.

Vauquois's Triangle



Approaches to MT

Classification based on the **computational architecture** of MT, also fuzzy:

- **Hand-crafted**: knowledge for analysis, transfer, generation, meaning representation, or direct translation is manually developed
  - most of commercial MT systems fall in this category
  - requires lots of human labor and expertise
  - includes: rule-based MT
- **Machine-learned**: representations are implemented by mathematical models learnable from data, e.g. parallel corpora of human translations
  - much less human effort is needed
  - requires huge amounts of data, the more, the better!
  - includes: statistical MT and example-based MT

Interlingua-Based Translation (C-STAR, 1999)

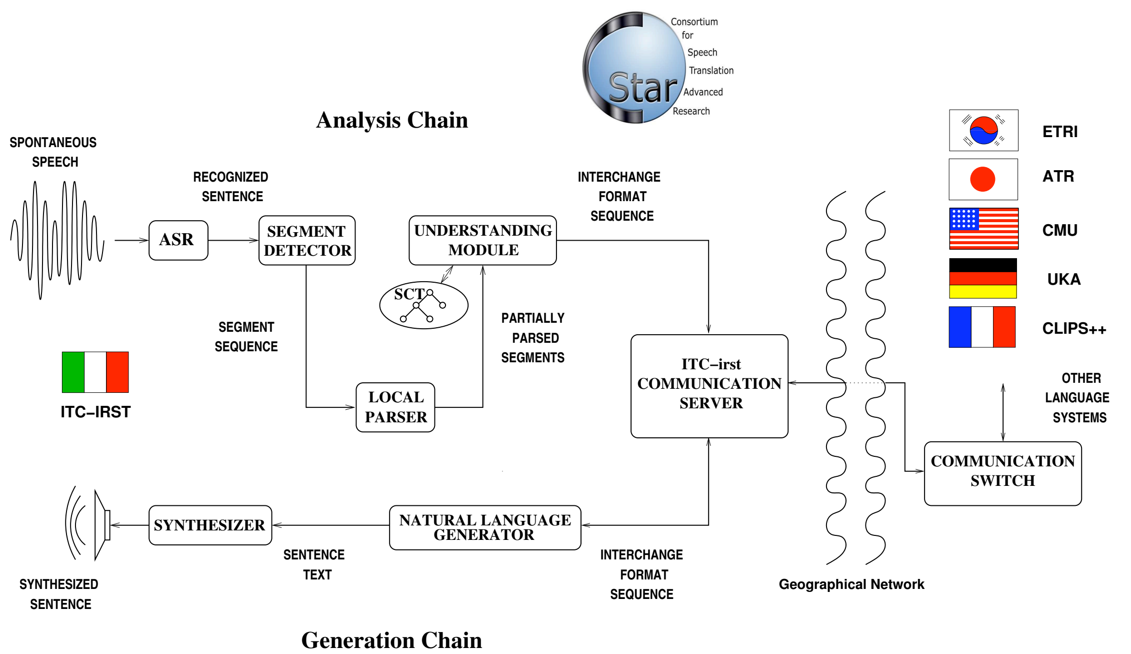
- S[1] : I'm arriving on june sixth
- I: `give-information+temporal+arrival (who=I, time=(june, md6))`
- T: my arrival time is sixth of june

- S: no that's not necessary
- I: `negate`
- T: no

- S: and i was wondering what you have in the way of rooms available during that time
- I: `request-information+availability+room (room-type=question)`
- T: what kind of rooms are available?

[1] S: speech (English), I: Interlingua, T: translation (English)

Interlingua-Based Translation (C-STAR, 1999)



Example-Based Translation

- **Assumption: people translate by analogy**
  - Decompose a sentence into phrases
  - Translate phrases by analogy to previous translations
  - Properly compose translation fragments into one long sentence
- **Given a parallel corpus of translation examples**

English	Japanese
How much is that red umbrella?	Ano akai kasa wa ikura desu ka.
How much is that small camera?	Ano chiisai kamera wa ikura desu ka.

- **Learn Translation patterns**

How much is that X?	$\longrightarrow$	Ano X wa ikura desu ka.
red umbrella	$\longrightarrow$	akai kasa
small camera?	$\longrightarrow$	chiisai kamera

Classical SMT Framework

Let $\mathbf{f}$ be any text in the source (foreign) language. The most probable translation $\hat{\mathbf{e}}$ is searched among texts in the target (English) language through the following statistical decision criterion:

$$\hat{\mathbf{e}} = \arg \max_{\mathbf{e}} \Pr(\mathbf{f} \mid \mathbf{e}) \Pr(\mathbf{e}) \quad (1)$$

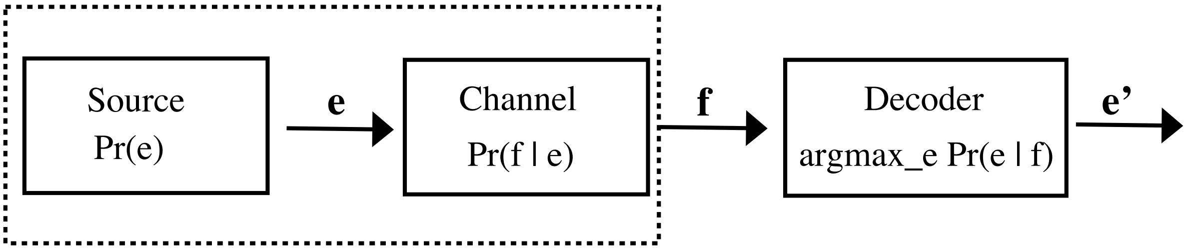
The **computational problems of SMT**:

- language modeling: estimating probabilities $\Pr(\mathbf{e})$
- translation modeling: estimating probabilities $\Pr(\mathbf{f} \mid \mathbf{e})$
- search problem: carrying out the optimization criterion (1)

Remarks

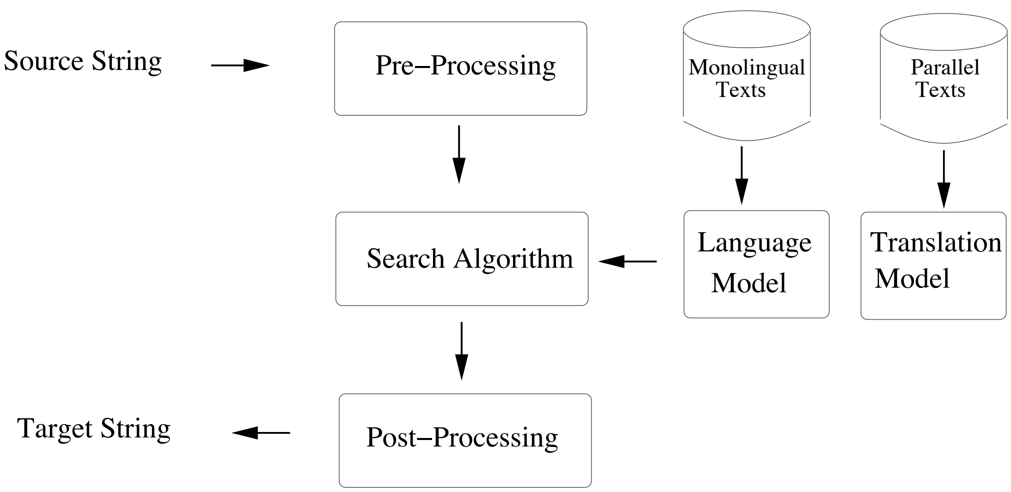
- all translation pairs are plausible, in principle, but have different probs
- although theory is presented with target English it is general

Noisy Channel Model of SMT



- English strings are generated by an unknown source and translated by an unknown channel into the foreign language
- MT translation is the task of decoding a foreign language into English
- Solution: estimate source and channel models and compute the decoding step.

Classical SMT Architecture



Example of Parallel Corpus

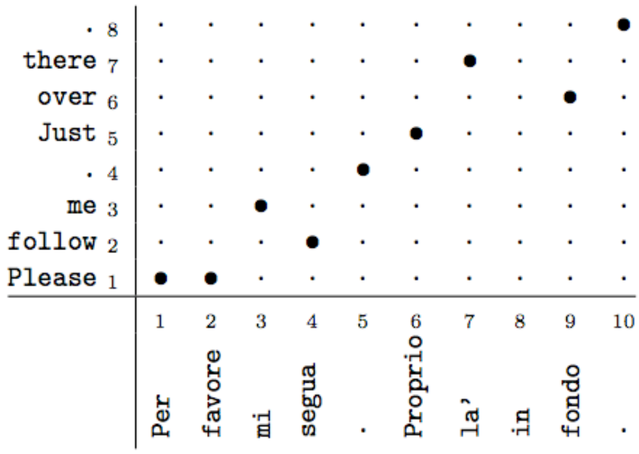
Darum liegt die Verantwortung für das Erreichen des Effizienzzieles und der damit einhergehenden CO2 -Reduzierung bei der Gemeinschaft , die nämlich dann tätig wird , wenn das Ziel besser durch gemeinschaftliche Massnahmen erreicht werden kann . Und genaugenommen steht hier die Glaubwürdigkeit der EU auf dem Spiel .

That is why the responsibility for achieving the efficiency target and at the same time reducing CO2 lies with the Community , which in fact takes action when an objective can be achieved more effectively by Community measures . Strictly speaking , it is the credibility of the EU that is at stake here .

Notice different positions of corresponding verb groups. **MT has to take into account word re-ordering!**

Translation Model and Alignments

- Translation has to consider possible alignments $\mathbf{a}$ between words in $\mathbf{f}$ and $\mathbf{e}$.
- Formally, alignments $\mathbf{a}$ are maps from positions of $\mathbf{f}$ to positions of $\mathbf{e}$.
- These and even more general alignments are learnable from translation examples.
- Notice, alignments induce word re-ordering



IBM Word-based Translation

Search Criterion

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{f} \mid \mathbf{e}) \Pr(\mathbf{e}) = \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \Pr(\mathbf{e})$$
$$\approx \arg\max_{\mathbf{e}} \max_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \ \Pr(\mathbf{e})$$

The max approximation is taken for computational reasons, to avoid summing over all alignments $\mathbf{a}$ from $\mathbf{f}$ to $\mathbf{e}$

- Alignment Model $Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$ is factorized into:
  - $p_\gamma(\mathbf{a} \mid \mathbf{e}) \rightarrow$ reordering model
  - $p_\theta(\mathbf{f} \mid \mathbf{e}, \mathbf{a}) \rightarrow$ lexicon model
- Language Model is
  - $Pr(\mathbf{e}) = p_\mu(\mathbf{e}) \rightarrow$ an n-gram language model

Alignment Model

An exact probability decomposition of $\Pr(\mathbf{f} = f_1^m, \mathbf{a} = a_1^m \mid \mathbf{e} = e_1^l)$ also takes into account the lengths of $\mathbf{e}$ and $\mathbf{f}$:

$$\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \Pr(m \mid \mathbf{e}) \Pr(\mathbf{a} \mid \mathbf{e}, m) \Pr(\mathbf{f}, \mid \mathbf{e}, \mathbf{a}, m)$$

Example of approximations which factorize further:

- Length Model: $$\Pr(m \mid \mathbf{e}) \approx p(m \mid l)$$
- Re-ordering model: $$\Pr(\mathbf{a} \mid \mathbf{e}, m) \approx \prod_{j=1}^{m} p(a_j \mid a_{j-1})$$
- Lexicon model: $$\Pr(\mathbf{f} \mid \mathbf{e}, \mathbf{a}, m) \approx \prod_{j=1}^{m} p(f_j \mid e_{a_j})$$

Language Model: Smoothed N-gramsl

The purpose of LMs is to compute the probability $\Pr(\mathbf{e}_1^T)$ of any sequence of words $\mathbf{e}_1^T = e_1 \ldots, e_t, \ldots, e_T$.

- **N-gram LMs** use the approximation:

$$\Pr(\mathbf{e}_1^T) \approx P(e_1) \prod_{t=2}^{T} \Pr(e_t \mid e_{t-n+1} \ldots e_{t-1})$$

  i.e. limit dependence to previous n-1 words
- single probs are computed by smoothing relative frequencies of n-grams collected on a huge text sample in the target language.
- the LM probability can be computed incrementally on the target string.

Search Problem: Decoding Algorithm

Given a statistical alignment model, a language model, and a source sentence, the task of the search procedure is to find the most likely translation:

$$\hat{\mathbf{e}} = \arg \max_{\mathbf{e}} p(\mathbf{e}) \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

Generally the **maximum approximation** is applied:

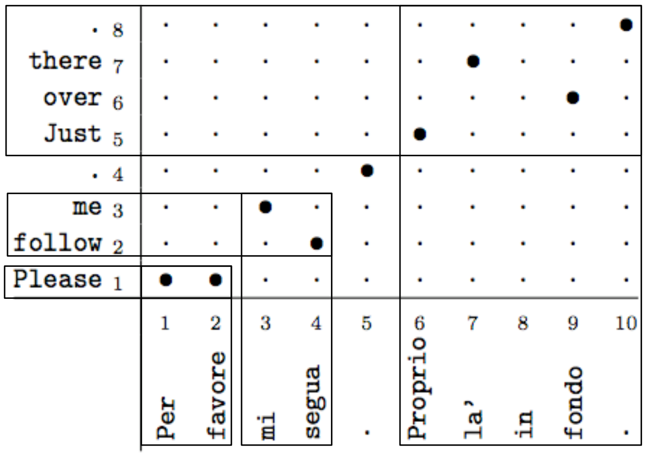
$$\hat{\mathbf{e}} = \arg \max_{\mathbf{e}} p(\mathbf{e}) \max_{\mathbf{a}} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

Complexity of the decoding problem mainly depends on word-reordering:

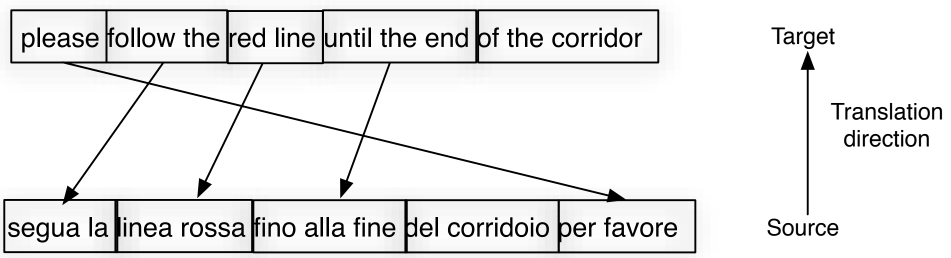
- monotonic translation (no word re-ordering): polynomial
- local reordering: high-polynomial
- all permutations: NP-hard

Alignments and Phrases

- Word alignments can be used to extract so called **phrase-pairs**
  - per favore - please
  - proprio la' in fondo . - just over there .
  - mi segua - follow me



From Word-based to Phrase-based Translation



- Phrases are finite sequence of words: n-grams with no linguistic meaning
- Source text is segmented into phrases
- Source phrases can be translated in different order
  - e.g. `per favore` is translated first!
- Source phrases are translated into target phrase

Search is performed bottom up by exploring a large number of hypotheses, taking into account possible segmentations, phrase re-orderings, translation alternatives.

The State of the Art

- **SMT is now a very competitive technology**
  – in many evaluations SMT outperformed rule-based MT
  – commercial systems perform likely better when not enough data are available
- Interest in SMT revamped around **seminal work at IBM in early 90'**
  – indeed the whole thing was started by **Warren Weaver in 1949**
- **Best performing SMT systems** use either:
  – brute force direct translation exploiting huge amounts of data
  – combination of direct translation and syntax-driven models
- **Automatic evaluation** metrics have dramatically boosted research in SMT:
  – model training directly optimizes the evaluation metric
- Several **evaluation campaigns** are organized every year:
  – NIST: news texts - Chi/Ara to Eng (2002-)
  – IWSLT: travelers speech - Chi/Jap/Ara/Ita to Eng (2004-)
  – TC-STAR: political speeches Spa-Eng, radio news Chi-Eng (2005-2007)

Evaluating MT Performance

How do we evaluate the output of a MT system?

- **Human MT evaluation**:
  – criteria: adequacy, fidelity, and fluency
  – pros: very accurate, high quality
  – cons: expensive and slow
- **Automatic MT evaluation**:
  – criteria: "similarity" to professional human translation
  – pros: inexpensive and quick
  – cons: quality is "slightly" lower than human check

Evaluation bottleneck: MT developers need to monitor the effect of daily changes to their systems in order to weed out bad ideas from good ideas!

Automatic Evaluation of MT

Automatic scoring methods typically compare the output against multiple high-quality human translations, called references:

- **Word alignment methods**
  - WER: ratio of smallest edit distance and output length
  - SER: 0 if WER is 0, and 1 otherwise
- **N-gram matching methods**
  - BLEU: compute weighted sum of counts of the matching n-grams
  - NIST: modification of BLEU
- **Task completion methods**
  - CLIR: compare IR performance with automatic and manual translations
  - IE: check information extraction performance
  - others

Example 1: Arabic English

Human	Dubai 2 - 7 (AFP) - The Secretary-General of the United Nations Kofi Annan said he would donate the international Zayed Prize for the Environment , which he received on Monday night in Dubai worth 500000 dollars , to setup a foundation for agriculture and educating girls in Africa .
Machine	Dubai 2-7 (AFP) - United Nations Secretary-General Kofi Annan said that the award will Zayed International Environment, which received Monday evening in Dubai worth 500,000 dollars to establish an institution for agriculture and education of girls in the African continent.

Example 1: Arabic English

Human	Dubai 2 - 7 (AFP) - The Secretary-General of the United Nations Kofi Annan said he would donate the international Zayed Prize for the Environment , which he received on Monday night in Dubai worth 500000 dollars , to setup a foundation for agriculture and educating girls in Africa .
Machine	Dubai 2-7 (AFP) - United Nations Secretary-General Kofi Annan said that the award ~~will~~ Zayed International Environment, which ... he ... received ... on... Monday evening in Dubai worth 500,000 dollars ... , will be donated ... to establish an institution for agriculture and education of girls in the African continent.

Example 2: Arabic English

Human	New York (The United Nations) 2 - 8 (AFP) - United Nations Secretary General Kofi Annan expressed his concern today , Tuesday , about the wave of targeted liquidations being carried out by Israel in Gaza and the West Bank , and he also condemned the rocket attacks targeting the Hebrew State , according to his spokesman .
Machine	New York (United Nations) 2-8 (AFP) - United Nations Secretary General Kofi Annan expressed concern today, Tuesday, the wave of qualifiers quality by Israel in Gaza and the West Bank, also condemned the missile attacks against the Jewish state, his spokesman said.

Example 2: Arabic English

Human	New York (The United Nations) 2 - 8 (AFP) - United Nations Secretary General Kofi Annan expressed his concern today , Tuesday , about the wave of targeted liquidations being carried out by Israel in Gaza and the West Bank , and he also condemned the rocket attacks targeting the Hebrew State , according to his spokesman .
Machine	New York (United Nations) 2-8 (AFP) - United Nations Secretary General Kofi Annan expressed concern today, Tuesday, ... about ... the wave of ~~qualifiers quality~~ targeted liquidations by Israel in Gaza and the West Bank, ... and he ... also condemned the missile attacks against the Jewish state, his spokesman said.

Example 3: Chinese English

Human	Today was the Catholic Church's annual " Life Day " . Pope Benedict XVI delivered a speech in St . Peter's Basilica , in which he criticized that the hedonism of wealthy society impairs the Christian value system of respect for life , and he strongly condemned abortion and euthanasia .
Machine	Today is the "life" of the Catholic Church once a year, when 16 of the pope delivered a speech in St. Peter's cathedral, criticized the joy of an affluent society, undermine the values of the Christian faith to respect life, and strongly condemned euthanasia and abortion.

Example 3: Chinese English

Human (?)	Today was the Catholic Church's annual " Life Day " . Pope Benedict XVI delivered a speech in St . Peter's Basilica , in which he criticized ~~that~~ the hedonism of ...our... wealthy society ...which... impairs the Christian value system of respect for life , and he strongly condemned abortion and euthanasia .
Machine	Today is the "life ..day..." of the Catholic Church ~~once a year~~, ~~when 16 of~~ the pope delivered a speech in St. Peter's cathedral, ...he... criticized the joy of an affluent society, ... that... undermines the values of the Christian faith to respect life, and strongly condemned euthanasia and abortion.

Example 4: Chinese English

Human	The Pope told thousands of believers making the pilgrimage to St . Peter's Basilica , " Life is often glorified during times of happiness , but no longer respected during times of sickness and trouble or when it is impaired . "
Machine	The pope told thousands who came to St. Peter's church followers, "when the joys of life were often, but sick or disabled, will no longer be respected."

Example 4: Chinese English

Human	The Pope told thousands of believers making the pilgrimage to St . Peter's Basilica , " Life is often glorified during times of happiness , but no longer respected during times of sickness and trouble or when it is impaired . "
Machine	The pope told thousands ... of followers... who came to St. Peter's church ~~followers~~, "when the~~-~~re is joys ~~of~~ life ~~were~~..was.. often ..glorified.., but ...when... sick or disabled, ~~will~~..it is.. no longer ~~be~~ respected."