New Paradigms for Machine Translation

Carnegie Mellon University

Jaime Carbonell et al

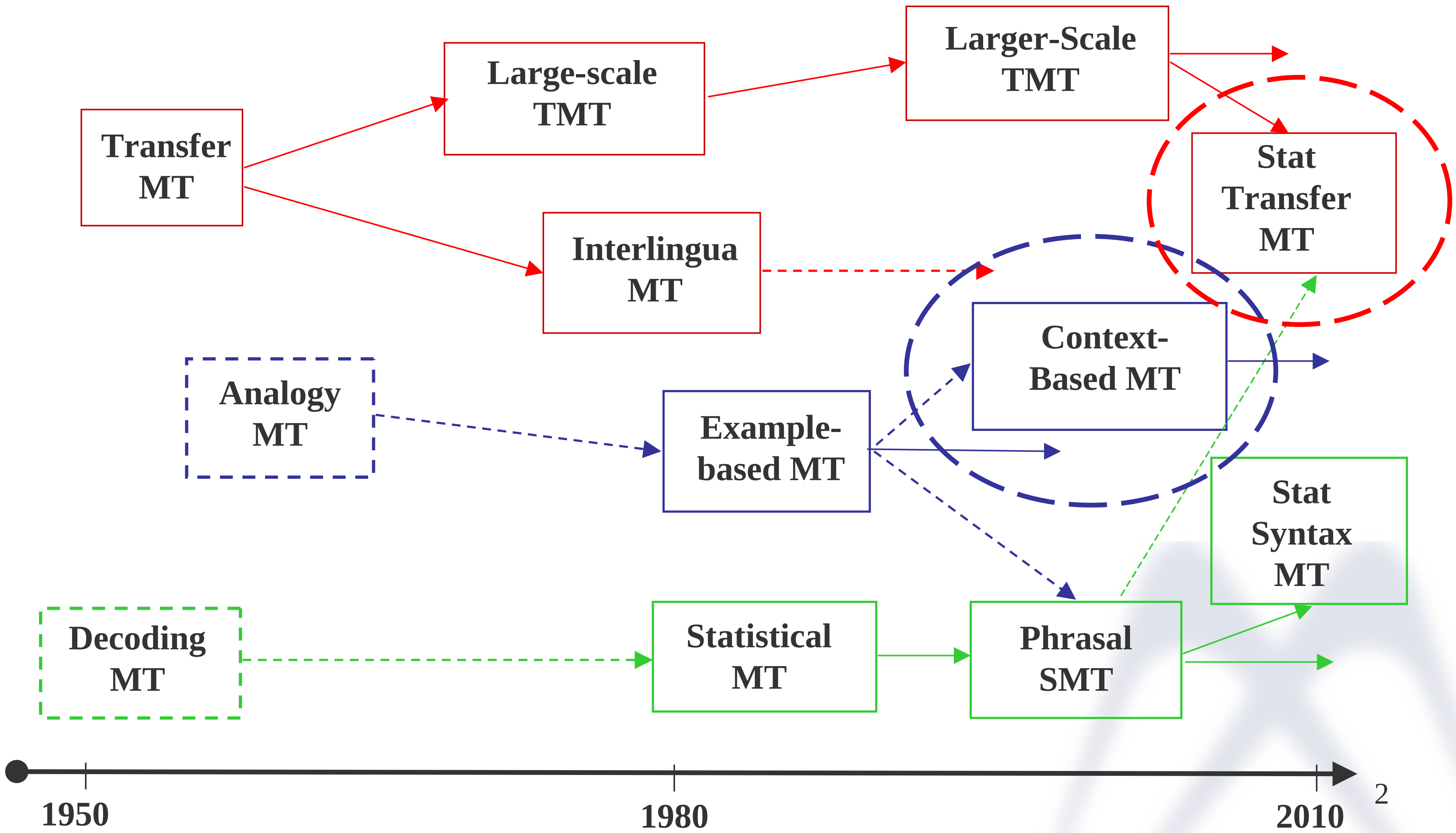
Context-Based MT

1. Pure unsupervised learning
2. Monolingual text only
3. Evaluations and Examples
4. Detecting & Exploiting Synonymy

Statistical Transfer

1. Learning transfer rules
2. Inducing tree alignments
3. Long-distance re-ordering

An Evolutionary Tree of MT Paradigms



Context Needed to Resolve Ambiguity

Example: English → Japanese

Power **line** – densen (電線)
Subway **line** – chikatetsu (地下鉄)
(Be) on **line** – onrain (オンライン)
(Be) on the **line** – denwachuu (電話中)
Line up – narabu (並ぶ)
Line one's pockets – kanemochi ni naru (金持ちになる)
Line one's jacket – uwagi o nijuu ni suru (上着を二重にする)
Actor's **line** – serifu (セリフ)
Get a **line** on – joho o eru (情報を得る)

Sometimes local context suffices (as above) → n-grams help
. . . but sometimes not

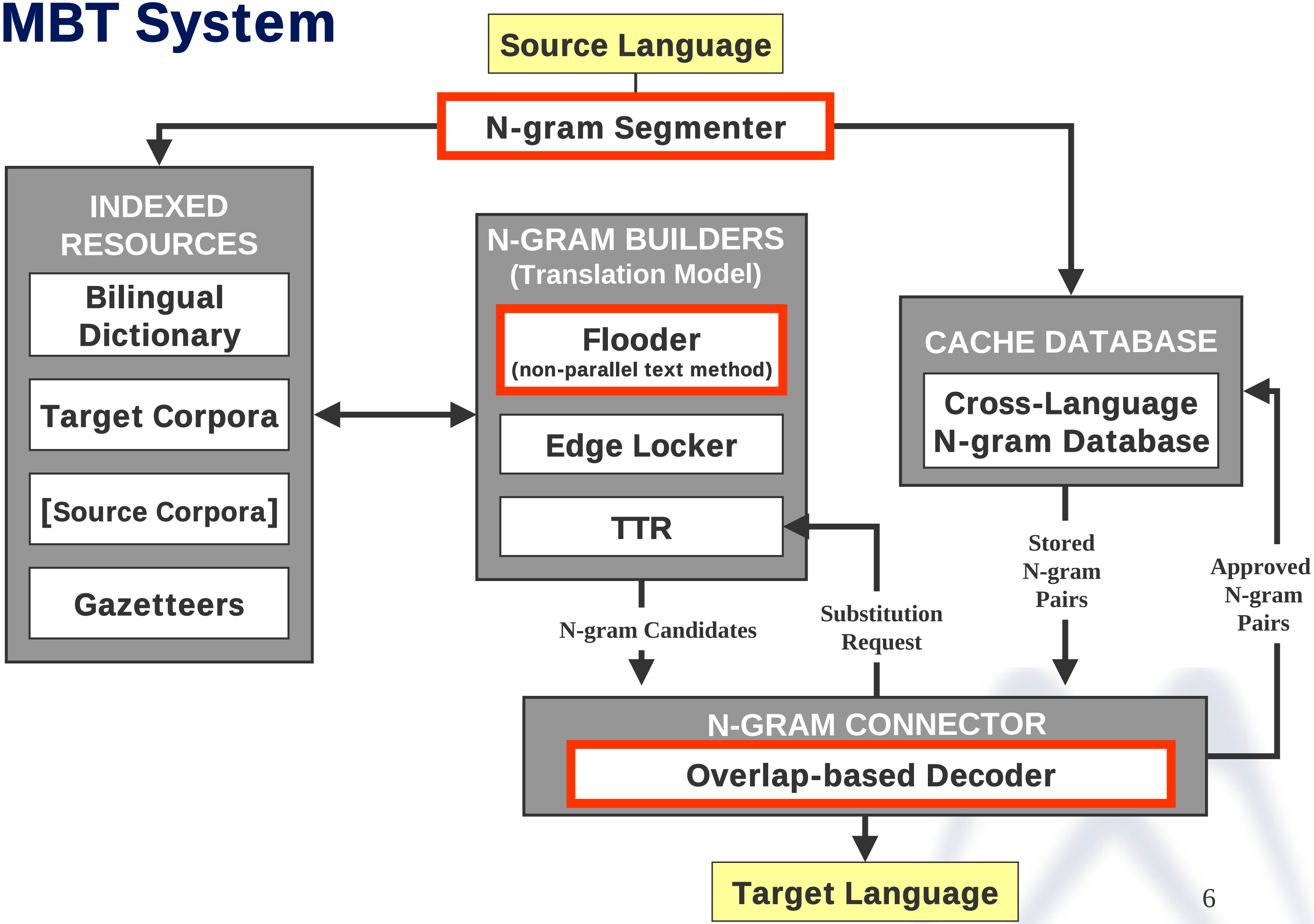
CONTEXT: More is Better

- **Examples requiring longer-range context:**
  - *"The **line** for the new play **extended for 3 blocks**."*
  - *"The **line** for the new play was changed by the **scriptwriter**."*
  - *"The **line** for the new play got **tangled with the other props**."*
  - *"The **line** for the **new play** better protected the **quarterback**."*
- **CBMT approach:**
  - Translation model uses 7-to-10 grams (+ 2 w's left, 2 right)
  - Overlap decoder cascades context throughout sentence
  - Also permits greater lexical reordering (e.g., for Chinese-English)

Parallel Text: Requiring Less is Better (Requiring None is Best ☺)

- **Challenge**
  - There is just not enough to approach human-quality MT for major language pairs (we need ~100X to ~10,000X)
  - Much parallel text is not on-point (not on domain)
  - Rare languages or distant pairs have very little parallel text
- **CBMT Approach** [Abir, Carbonell, Sofizade, …]
  - *Requires* ***no parallel text****, no transfer rules . . .*
  - *Instead, CBMT needs*
    - *A fully-inflected* ***bilingual dictionary***
    - *A (very large)* ***target-language-only corpus***
    - *A (modest)* ***source-language-only corpus*** *[optional, but preferred]*

CMBT System



Source Language
N-gram Segmenter
INDEXED RESOURCES
Bilingual Dictionary
Target Corpora
[Source Corpora]
Gazetteers
N-GRAM BUILDERS
(Translation Model)
Flooder
(non-parallel text method)
Edge Locker
TTR
CACHE DATABASE
Cross-Language N-gram Database
N-gram Candidates
Substitution Request
Stored N-gram Pairs
Approved N-gram Pairs
N-GRAM CONNECTOR
Overlap-based Decoder
Target Language

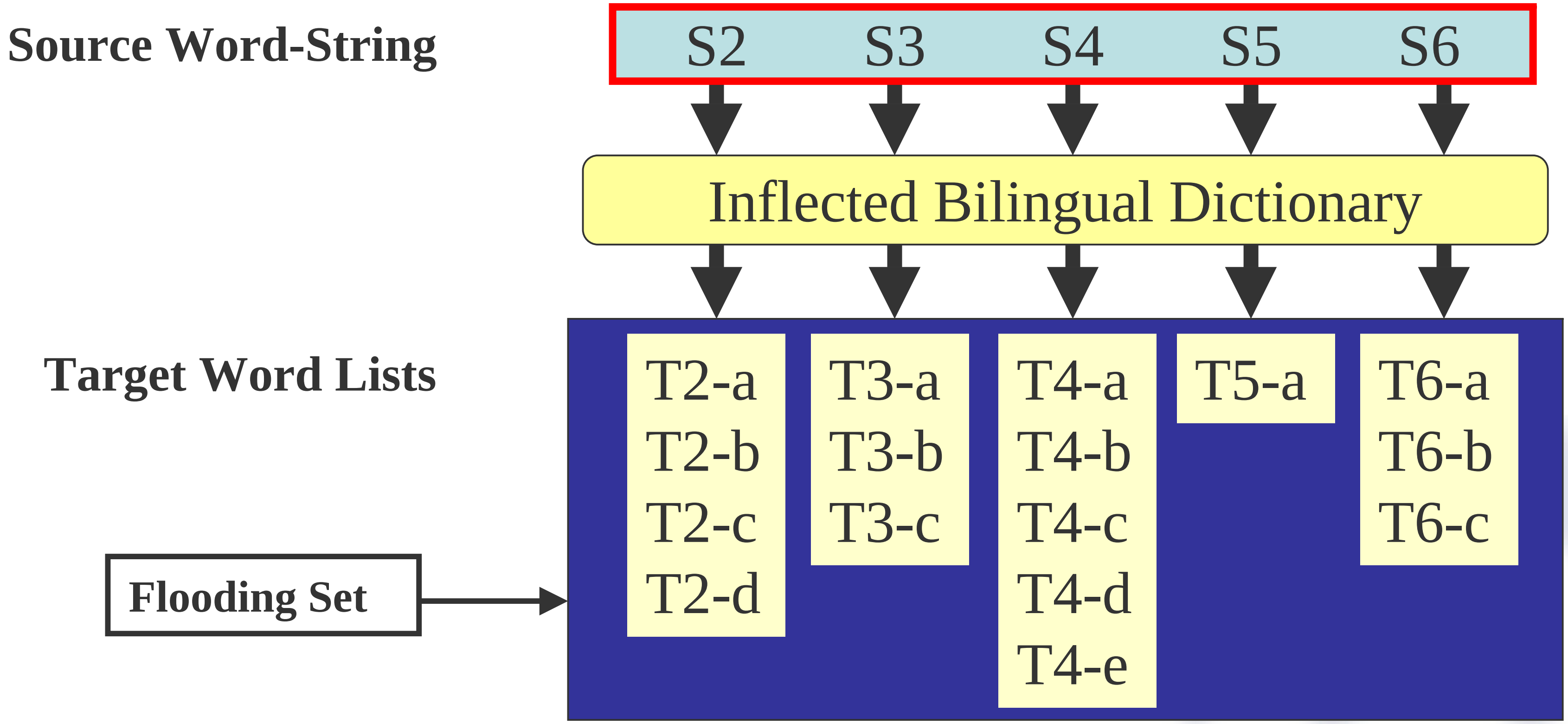
Step 1: Source Sentence Chunking

- Segment source sentence into overlapping n-grams via sliding window
- Typical n-gram length 4 to 9 terms
- Each term is a word or a known phrase
- Any sentence length (for BLEU test: ave-27; shortest-8; longest-66 words)

S1	S2	S3	S4	S5	S6	S7	S8	S9
S1	S2	S3	S4	S5				
	S2	S3	S4	S5	S6			
		S3	S4	S5	S6	S7		
			S4	S5	S6	S7	S8	
				S5	S6	S7	S8	S9

Step 2: Dictionary Lookup

- Using bilingual dictionary, list all possible target translations for each source word or phrase



Step 3: Search Target Text

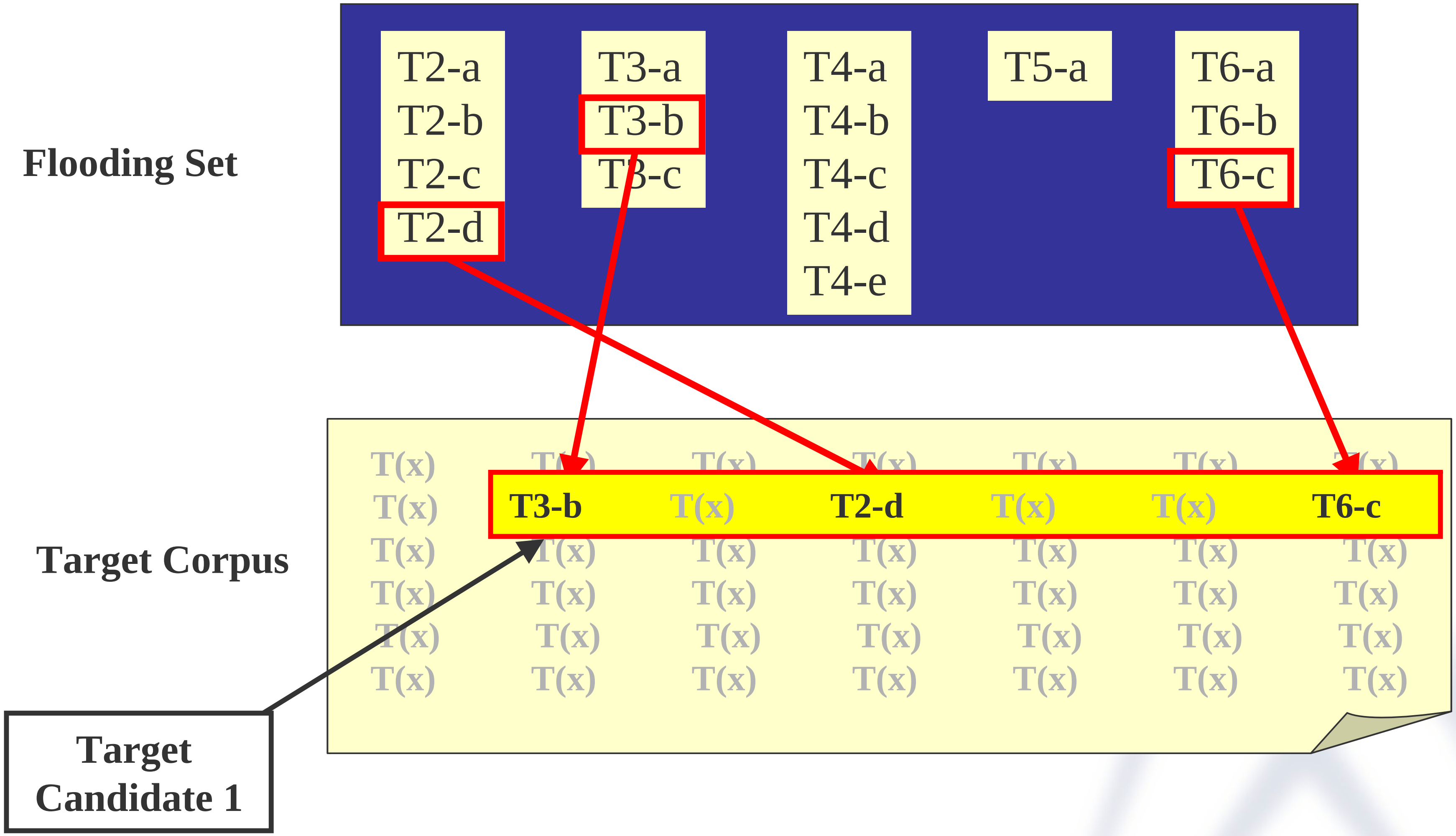
- Using the Flooding Set, search target text for word-strings containing one word from each group

Flooding Set

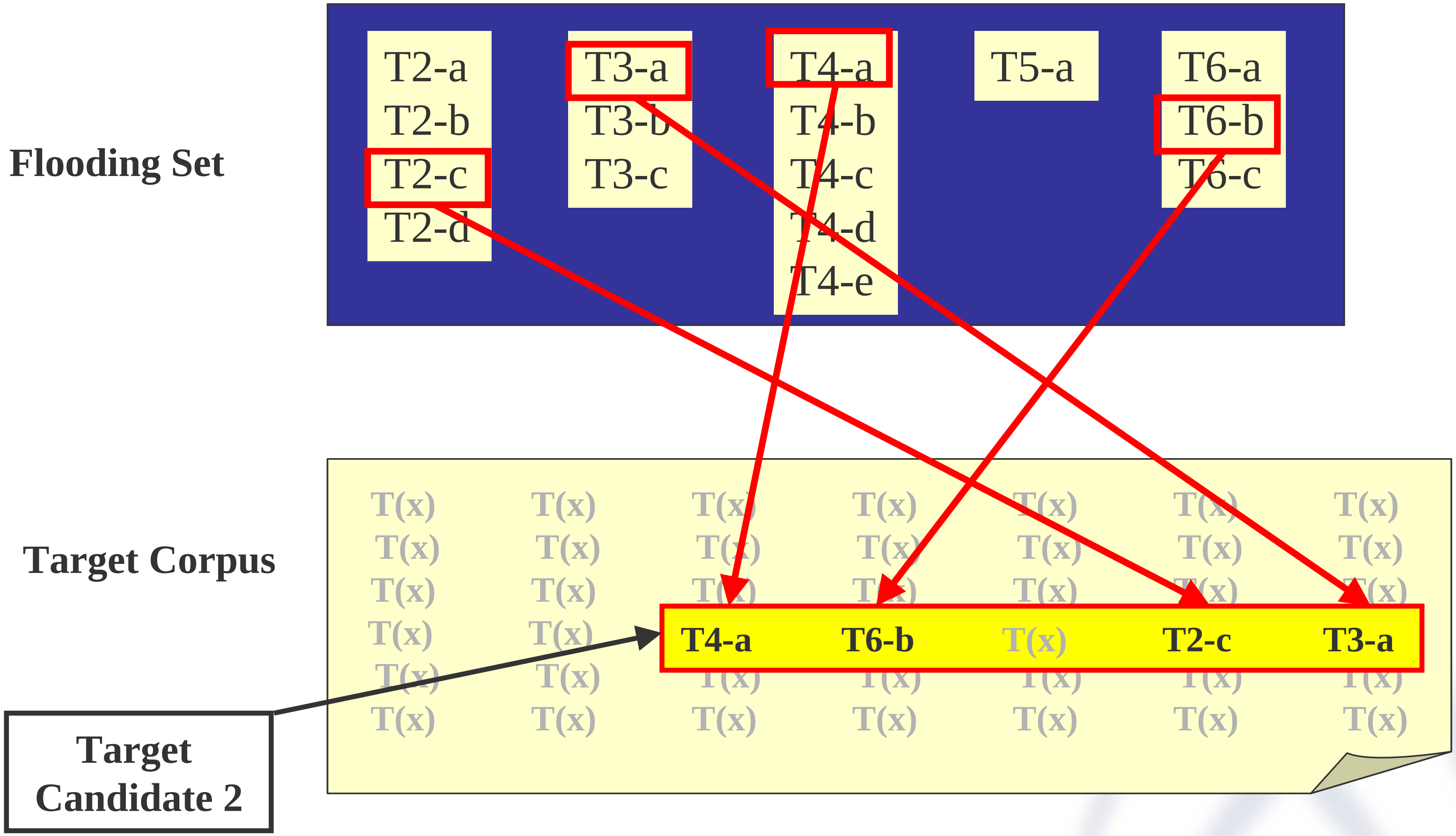
T2-a	T3-a	T4-a	T5-a	T6-a
T2-b	T3-b	T4-b		T6-b
T2-c	T3-c	T4-c		T6-c
T2-d		T4-d		
		T4-e		

- Find maximum number of words from Flooding Set in minimum length word-string
  - *Words or phrases can be in any order*
  - *Ignore function words in initial step (T5 is a function word in this example)*

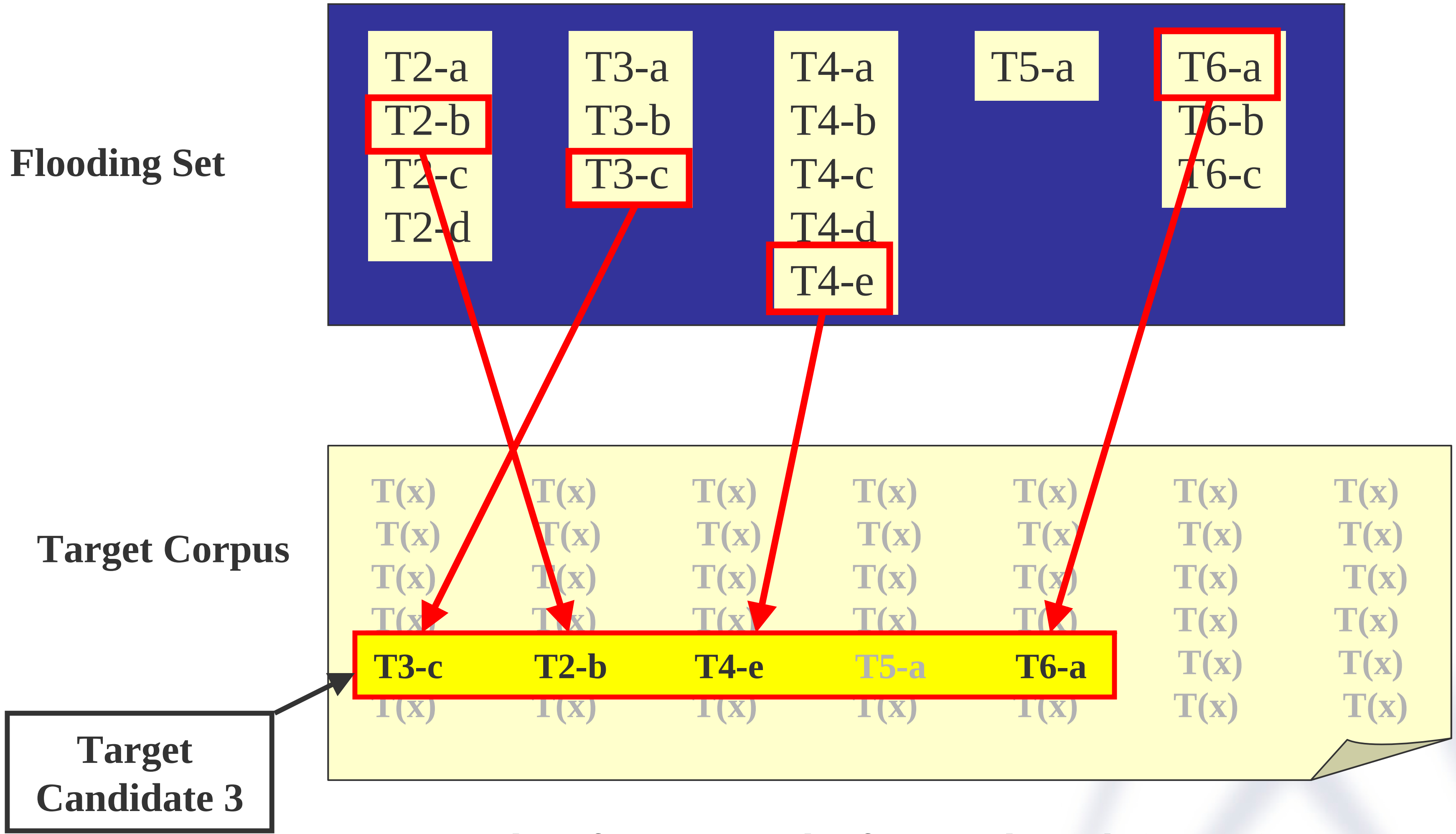
Step 3: Search Target Text (Example)



Step 3: Search Target Text (Example)



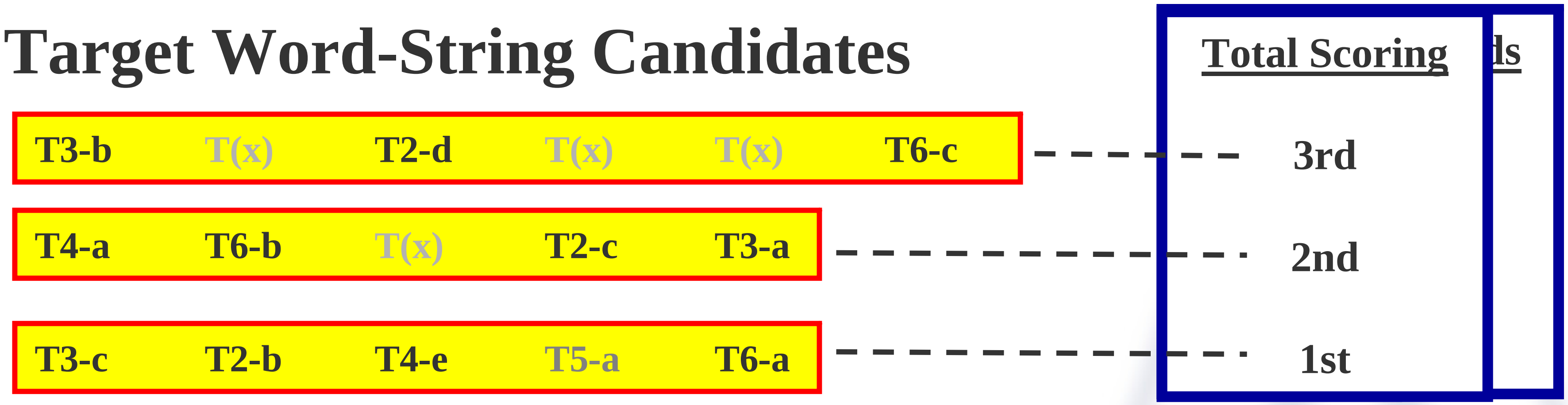
Step 3: Search Target Text (Example)



Reintroduce function words after initial match (e.g. T5)

Step 4: Score Word-String Candidates

- Scoring of candidates based on:
  - Proximity (minimize extraneous words in target n-gram ≈ precision)
  - Number of word matches (maximize coverage ≈ recall))
  - Regular words given more weight than function words
  - Combine results (e.g., optimize F_1 or p-norm or …)



Step 5: Select Candidates Using Overlap

(Propagate context over entire sentence)

Word-String 1 Candidates

T(x1)	T2-d	T3-c	T(x2)	T4-b	
T(x1)	T3-c	T2-b	T4-e		
T(x2)	T4-a	T6-b	T(x3)	T2-c	

Word-String 2 Candidates

T3-b	T(x3)	T2-d	T(x5)	T(x6)	T6-c
T4-a	T6-b	T(x3)	T2-c	T3-a	
T3-c	T2-b	T4-e	T5-a	T6-a	

Word-String 3 Candidates

T2-b	T4-e	T5-a	T6-a	T(x8)
T6-b	T(x11)	T2-c	T3-a	T(x9)
T6-b	T(x3)	T2-c	T3-a	T(x8)

Step 5: Select Candidates Using Overlap

Best translations selected via maximal overlap

Alternative 1

T(x2) T4-a T6-b T(x3) T2-c

T4-a T6-b T(x3) T2-c T3-a

T6-b T(x3) T2-c T3-a T(x8)

T(x2) T4-a T6-b T(x3) T2-c T3-a T(x8)

Alternative 2

T(x1) T3-c T2-b T4-e

T3-c T2-b T4-e T5-a T6-a

T2-b T4-e T5-a T6-a T(x8)

T(x1) T3-c T2-b T4-e T5-a T6-a T(x8)

A (Simple) Real Example of Overlap

Flooding → N-gram fidelity

Overlap → Long range fidelity

N-grams generated from Flooding

- **A United States soldier**
- **United States soldier died**
- **soldier died and two others**
- **died and two others were injured**
- **two others were injured Monday**

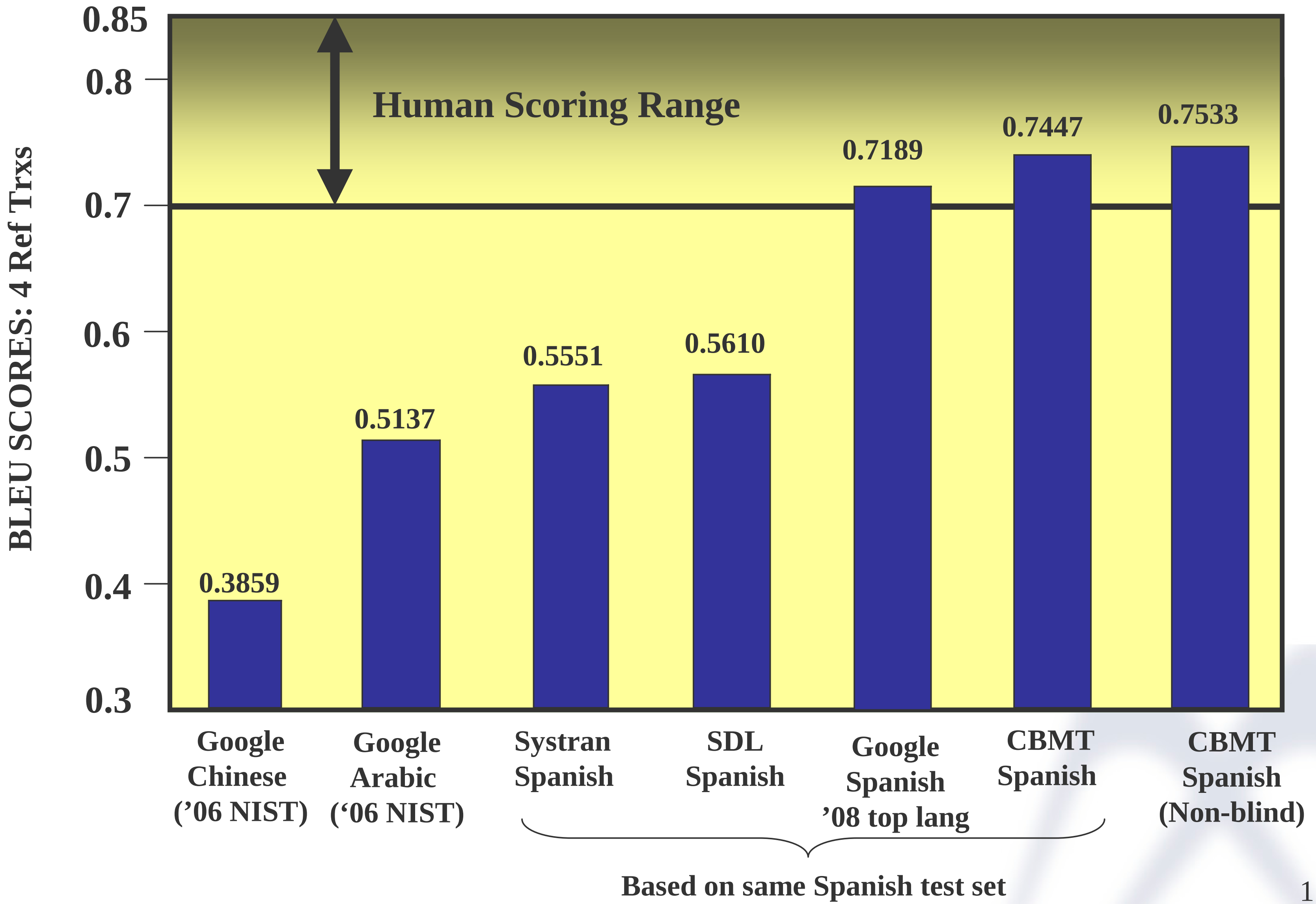
N-grams connected via Overlap

A United States soldier died and two others were injured Monday

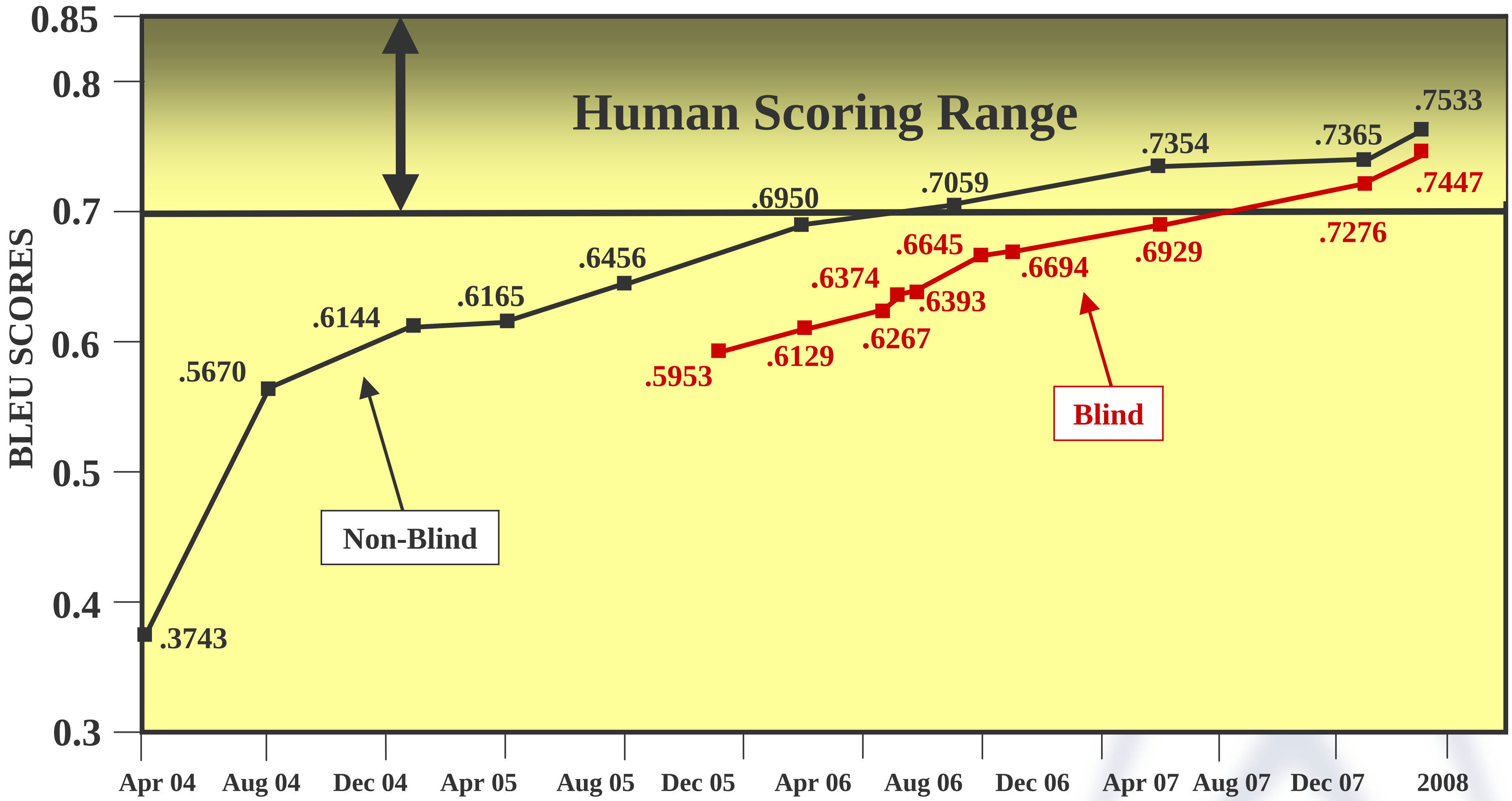
---

Systran: A soldier of the **wounded** United States died and other two were **east** Monday

System Scores



Historical CBMT Scoring



An Example

- Un soldado de Estados Unidos murió y otros dos resultaron heridos este lunes por el estallido de un artefacto explosivo improvisado en el centro de Bagdad, dijeron funcionarios militares estadounidenses

---

- **CBMT:** A United States soldier died and two others were injured monday by the explosion of an improvised explosive device in the heart of Baghdad, American military officials said.

- ***Systran*:** A soldier of the wounded United States died and other two were east Monday by the outbreak from an improvised explosive device in the center of Bagdad, said American military civil employees

BTW: Google's translation is identical to CBMT's

Beyond the Basics of CBMT

- What if a source word or phrase is not in the bilingual dictionary?
  - *Find near synonyms in source,*
  - *Replace and retranslate*
- What if overlap decoder fails to confirm any translation (e.g., insufficient target corpus)?
  - *Find near synonyms in target*
  - *Temporary token replacement (TTR)*

→**Need an automated near-synonym finder**

TTR Unsupervised Learning
Step 1: Document Search

- Search monolingual documents for occurrences of query.
- Each occurrence has a "signature" (words to left and right – together they form a "cradle").

Standard & Poor's indices are broad-based measures of changes in stock market conditions based on the performance of widely held common stocks . . . A large number of retirees are taking their money out of the stock market and putting it into safer money markets and fixed income investments . . . Funds across the board had their worst month in August but stabilized as the stock market rebounded for most of the summer . . . Measuring changes in stock market wealth have become a more important determinant of consumer confidence . . . PlanetWeb announced Friday that it would be de-listed from the NASDAQ stock market before the opening of trading on Tuesday . . . Some of these investors find it hard to exit troubled stock market and banking ventures . . . A direct correlation between money coming out of the stock market and money going into the bank do not exist . . . Users of the new system get results in real-time while sharing in the most extensive stock market information network available today . . .

TTR Unsupervised Learning
Step 2: Build Cradles

Left Signature	Middle	Right Signature
of changes in		conditions based on
out of the		and putting it
stabilized as the		rebounded for most
changes in		wealth have become
from the NASDAQ		before the opening
to exit troubled		and banking ventures
out of the		and money going
the most extensive		information network available

TTR Unsupervised Learning
Step 3: Fill Cradles with New Middle

Auto industry analysts have taken notice of changes in industry conditions based on reports from the major auto makers . . . Since the e-commerce bubble burst, the trend continues as investors are shifting capital out of the market and putting it into less volitile alternatives such as real estate despite liquidity limitations . . . Donations saw a dramatic drop in the first quarter but stabilized as the economy rebounded for most of the year . . . Investors simply “grin and bear it,” as roller-coaster changes in stock market wealth have become a commonplace occurrence . . . E-commerce pioneer WebPlanet received assurances from the NASDAQ stock exchange before the opening on Thursday that the stock would not be de-listed . . . Foreign parties who were interviewed noted that it was impossible to exit troubled federal government and banking ventures without an inside lobbying effort, oftentimes accompanied by a “consulting fee” . . . According to official Thai estimates, the relationship of money going out of the national market system and money going into the US stock market showed a strong correlation . . . The National Weather Center offers the most extensive government information network available, utilizing resources from every state weather agency . . .

TTR Unsupervised Learning
Step 3: Fill Cradles with New Middles

Left Signature	New Middle	Right Signature
of changes in	market	conditions based on
out of the	equities market	and putting it
changes in	market	wealth have become
stabilized as the	stock exchange	rebounded for most
from the NASDAQ	stock exchange	before the opening
out of the	national market	and money going
to exit troubled	major stock market	and banking ventures
the most extensive	government	information network available

TTR Unsupervised Learning
Step 4: Build Association List

Preliminary Association List for:
stock market

market (394)
stock exchange (292)
national market (189)
stock market® (85)
exchange (81)
equities market (61)
the stock market (48)
electronic exchange (32)
stocks exchange (30)

Scoring is a relative weight based on number of total occurrences and number of unique signatures that result appears in.

MM’s Association Builder

- Can generate lists of words and phrases that are synonymous to a query term or have other direct associations, such as class members or opposites.
- Can enhance search, text mining.

Term	Associations
terrorist organization	terrorist network / terrorist group / militant group / terror network extremist group / terrorist organisation / militant network
conference	meeting / symposium / convention / briefing / workshop
bin laden	bin ladin / bin-laden / osama bin laden / usama bin laden
nation’s largest	country’s largest / nation’s biggest / nation’s leading
watchful eye	direct supervision / close watch / stewardship / able leadership
it is safe to say	it’s fair to say / it is important to note / you will find / I can say it is important to recognize / it is well known / it is obvious

Examples of Alternative Spellings

Query

al qaeda	

Results
(partial)

al-qaida	(110)
al-qaeda	(109)
al-qaida	(24)
al-qa'eda	(5)
al queda	(4)
al- qaeda	(4)
al-qa'ida	(3)
al quaeda	(2)
al- qaida	(2)
al-quada	(1)

Other returns included: osama bin ladin (3), terrorist (3), international (3), islamic (2), worldwide (2), afghanistan-based (2) – among others

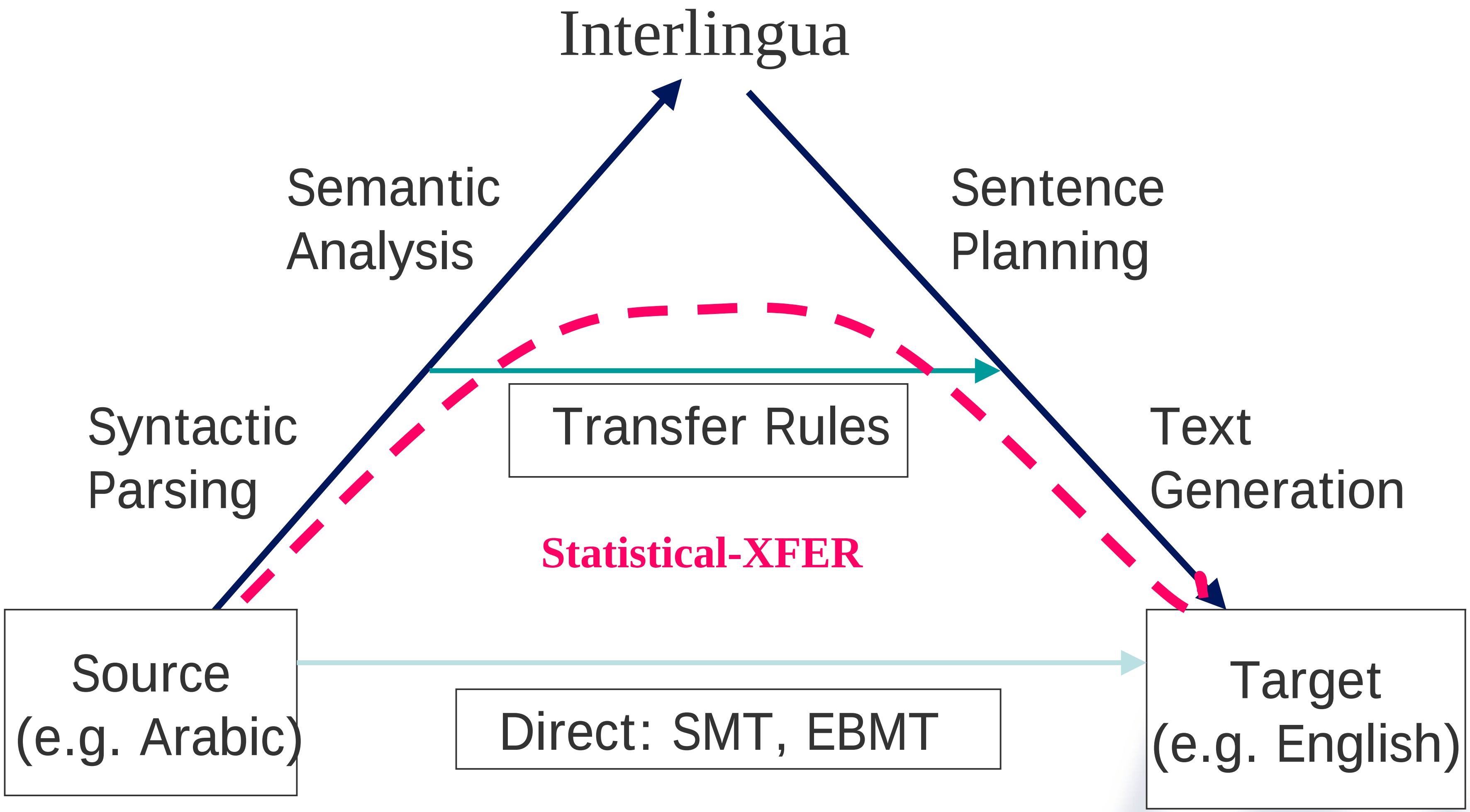
Stat-Transfer MT: Research Goals
(Lavie, Carbonell, Levin, Vogel & Students)

- Long-term research agenda (since 2000) focused on developing a unified framework for MT that addresses the core fundamental weaknesses of previous approaches:
  - ***Representation*** *– explore richer formalisms that can capture complex divergences between languages*
  - *Ability to handle* ***morphologically complex languages***
  - *Methods for* ***automatically acquiring MT resources*** *from available data and* ***combining them with manual resources***
  - *Ability to address both* ***rich and poor resource scenarios***
- Main research funding sources: NSF (AVENUE and LETRAS projects) and DARPA (GALE)

Stat-XFER: List of Ingredients

- **Framework:** Statistical search-based approach with syntactic translation transfer rules that can be acquired from data but also developed and extended by experts
- **SMT-Phrasal Base:** Automatic Word and Phrase translation lexicon acquisition from parallel data
- **Transfer-rule Learning**: apply ML-based methods to automatically acquire syntactic transfer rules for translation between the two languages
- **Elicitation**: use bilingual native informants to produce a small high-quality word-aligned bilingual corpus of translated phrases and sentences
- **Rule Refinement**: refine the acquired rules via a process of interaction with bilingual informants
- **XFER + Decoder**:
  - *XFER engine produces a lattice of possible transferred structures at all levels*
  - *Decoder searches and selects the best scoring combination*

Stat-XFER MT Approach

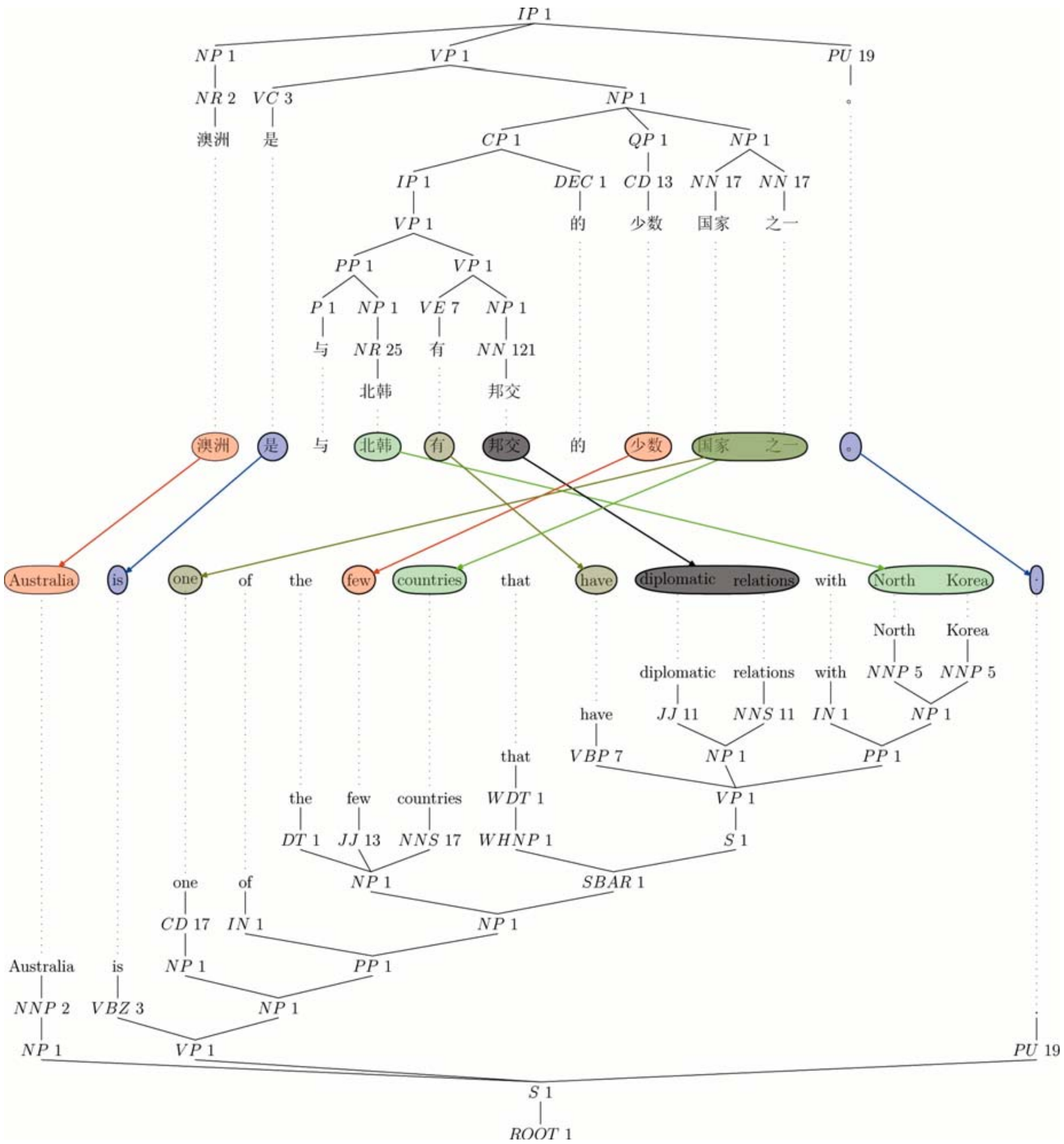


Syntax-driven Acquisition Process

Automatic Process for Extracting Syntax-driven Rules and Lexicons from sentence-parallel data:

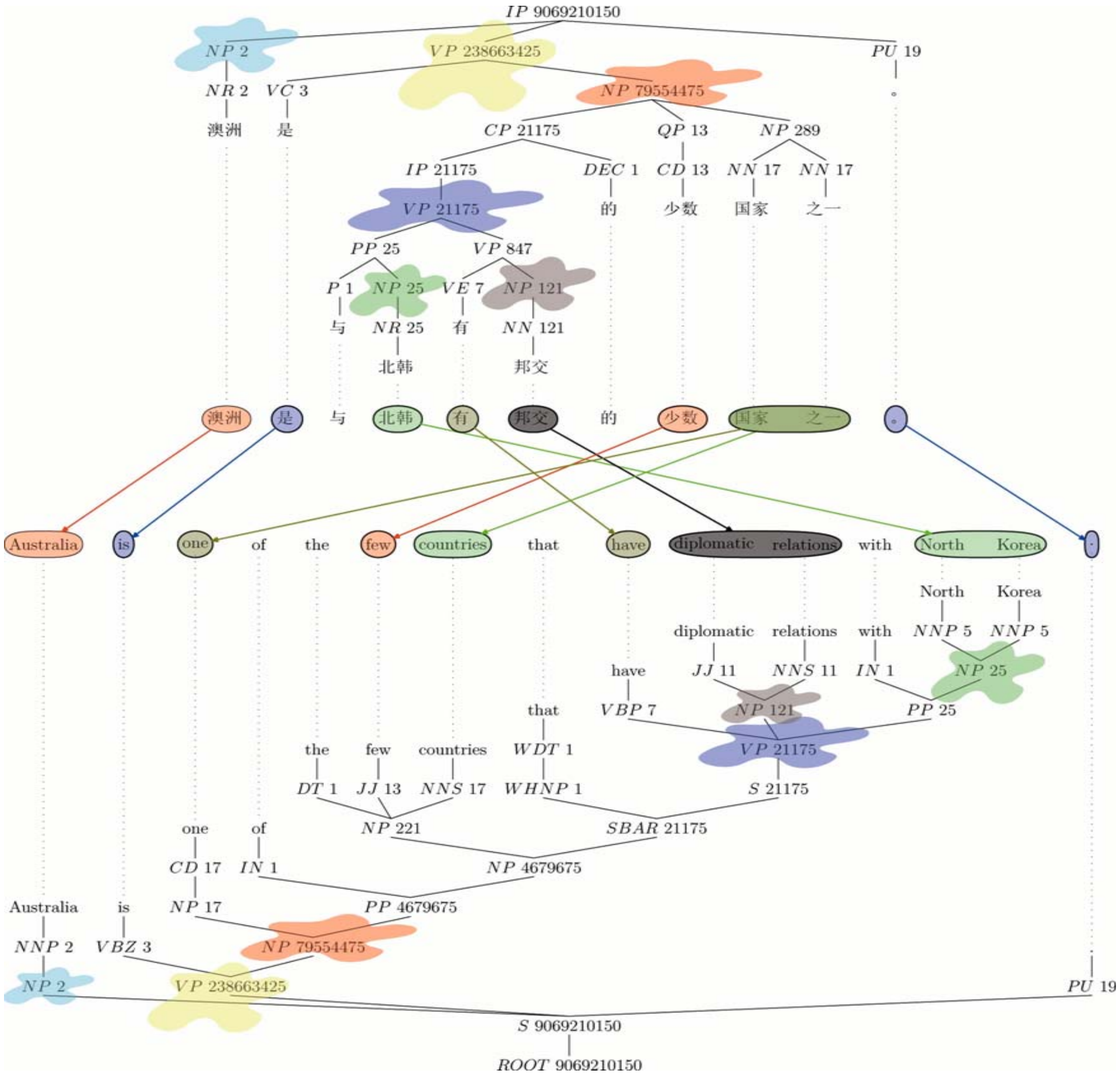
- ***Word-align*** *the parallel corpus (GIZA++)*
- ***Parse the sentences independently*** *for both languages*
- ***Tree-to-tree Constituent Alignment:***
  - ***Run our new Constituent Aligner*** *over the parsed sentence pairs*
  - ***Enhance alignments*** *with additional Constituent Projections*
- ***Extract all aligned constituents*** *from the parallel trees*
- ***Extract all derived synchronous transfer rules*** *from the constituent-aligned parallel trees*
- ***Construct a "data-base"*** *of all extracted parallel constituents and synchronous rules* ***with their frequencies*** *and model them statistically (assign them* ***relative-likelihood probabilities****)*

PFA Node Alignment Algorithm Example



- Any constituent or sub-constituent is a candidate for alignment
- Triggered by word/phrase alignments
- Tree Structures can be highly divergent

PFA Node Alignment Algorithm Example



- Tree-tree aligner enforces equivalence constraints and optimizes over terminal alignment scores (words/phrases)
- Resulting aligned nodes are highlighted in figure
- Transfer rules are partially lexicalized and read off tree.

Concluding Thoughts

- New/improved MT Paradigms are active areas for investigation
  - *Even for paradigmatic zealots: Why cannot transfer rules be automatically learned from data?*
  - *Why cannot we rely primarily on huge monolingual text for most of our action?*
- Caution 1: “Rigor engenders science, alas also mortis” – Herbert A. Simon (Nobel Laureate)
- Caution 2: There is a huge difference between a general theory & a system that respects it.
  - *Statistical decision theory + ML >> SMT*

Where will MT be in 4000 Years?