

New Paradigms for MT and IR

Carnegie Mellon University & Meaningful Machines

Jaime Carbonell

Machine Translation

1. Context-Based Machine Translation
2. New Paradigm for MT
3. Evaluations and Examples
4. Detecting & Exploiting Synonymy

Information Retrieval

1. Beyond mere relevance ranking
2. Clustering
3. Maximal Marginal Relevance
4. Personalized Search Agents

Language Technologies

SLOGAN

- “...right information”
- “...right people”
- “...right time”
- “...right medium”
- “...right language”
- “...right level of detail”

TECHNOLGY (e.g.)

- IR (search engines)
- Routing, personalization
- Anticipatory analysis
- Info extraction, speech
- Machine translation
- Summarization,
expansion

Exploiting Context for LT's

- Machine Translation
 - *Why* context is so necessary
 - *Context-Based MT* (new paradigm)
 - *Finding near-synonym phrases* (via context)
- Information Retrieval
 - *Context from clustering*
 - *Context from MMR-search*
 - *Personalized Search Profiles*

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)

... 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 4-to-10 grams (+ 2 w’s left, 2 right)
 - Overlap decoder cascades context throughout sentence
 - Permits greater lexical reordering (e.g., for Chinese-English)

Key Challenges for Corpus-Based MT

- **Long-Range Context**

- More is better
- How to incorporate it?
- How to do so tractably?
- “I conjecture that MT is AI complete” – H. Simon

- **State-of-the-Art**

- *Trigrams* → 4 or 5-grams in LM only (e.g., Google)
- *SMT* → *SMT+EBMT* (phrase tables, etc.)

- **Parallel Corpora**

- More is better
- Where to find enough of it?
- What about rare languages?
- “There’s no data like more data”
– IBM SMT circa 1992

- **State-of-the-Art**

- *Avoids rare languages (GALE only does Arabic & Chinese)*
- *Symbolic rule-learning from minimalist parallel corpus (AVENUE project at CMU)*

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

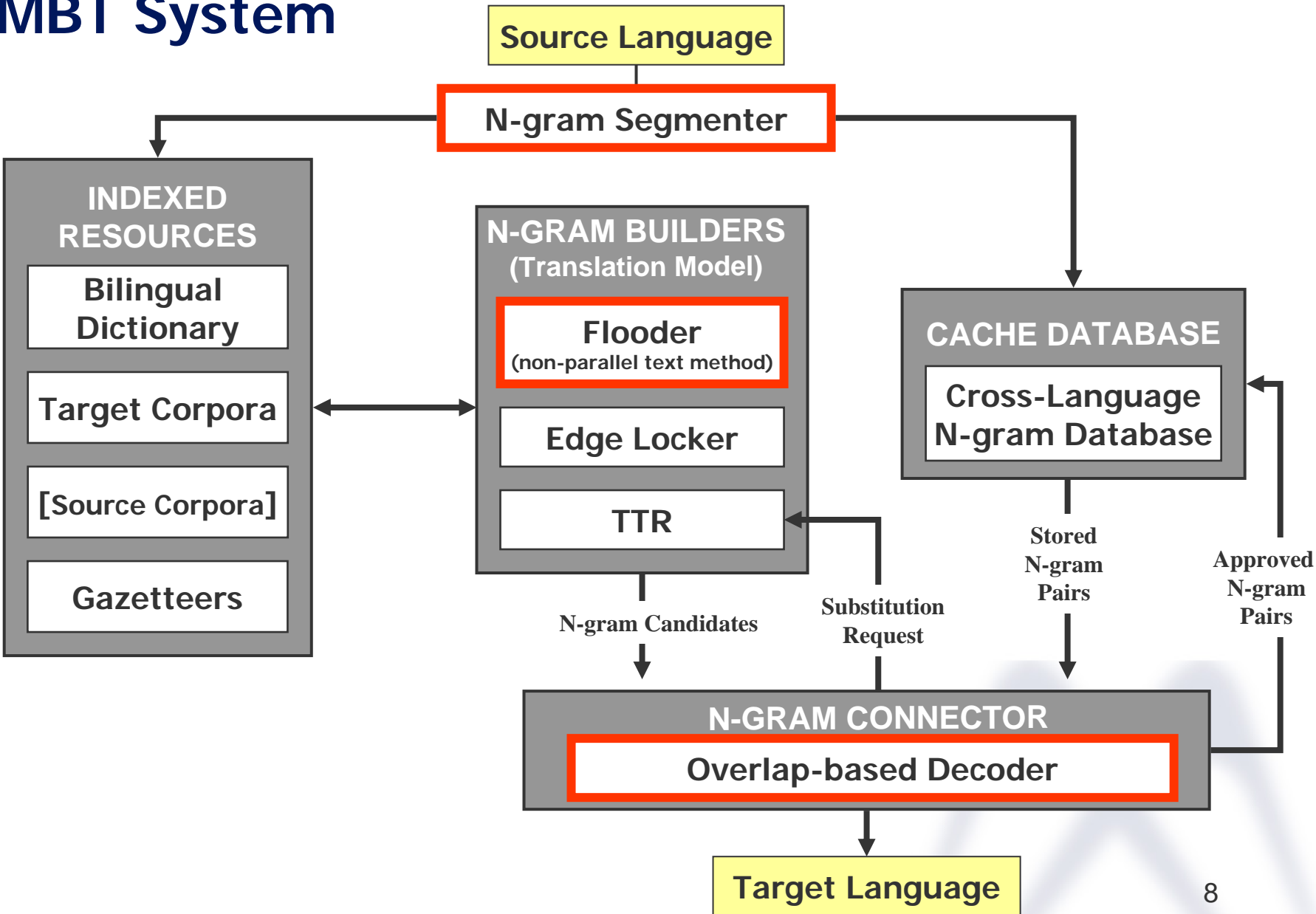
- **Challenge**

- There is just not enough to approach human-quality MT for major language pairs (we need ~1000X)
- Much parallel text is not on-point (not on domain)
- Rare languages or distant pairs have virtually no parallel text

- **CBMT Approach**

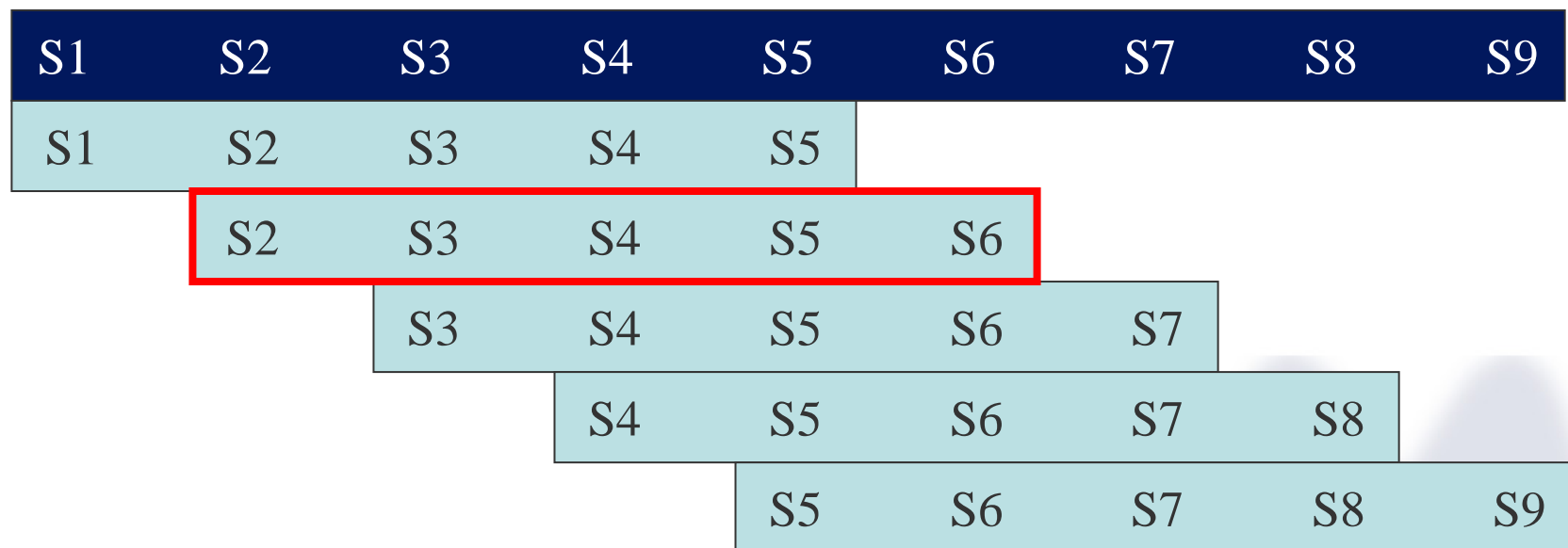
- *Requires no parallel text, no transfer rules . . .*
- *Instead, CBMT needs*
 - *A fully-inflected **bilingual dictionary***
 - *A (very large) **target-language-only corpus***
 - *A (modest) **source-language-only corpus** [optional, but preferred]*

CMBT System



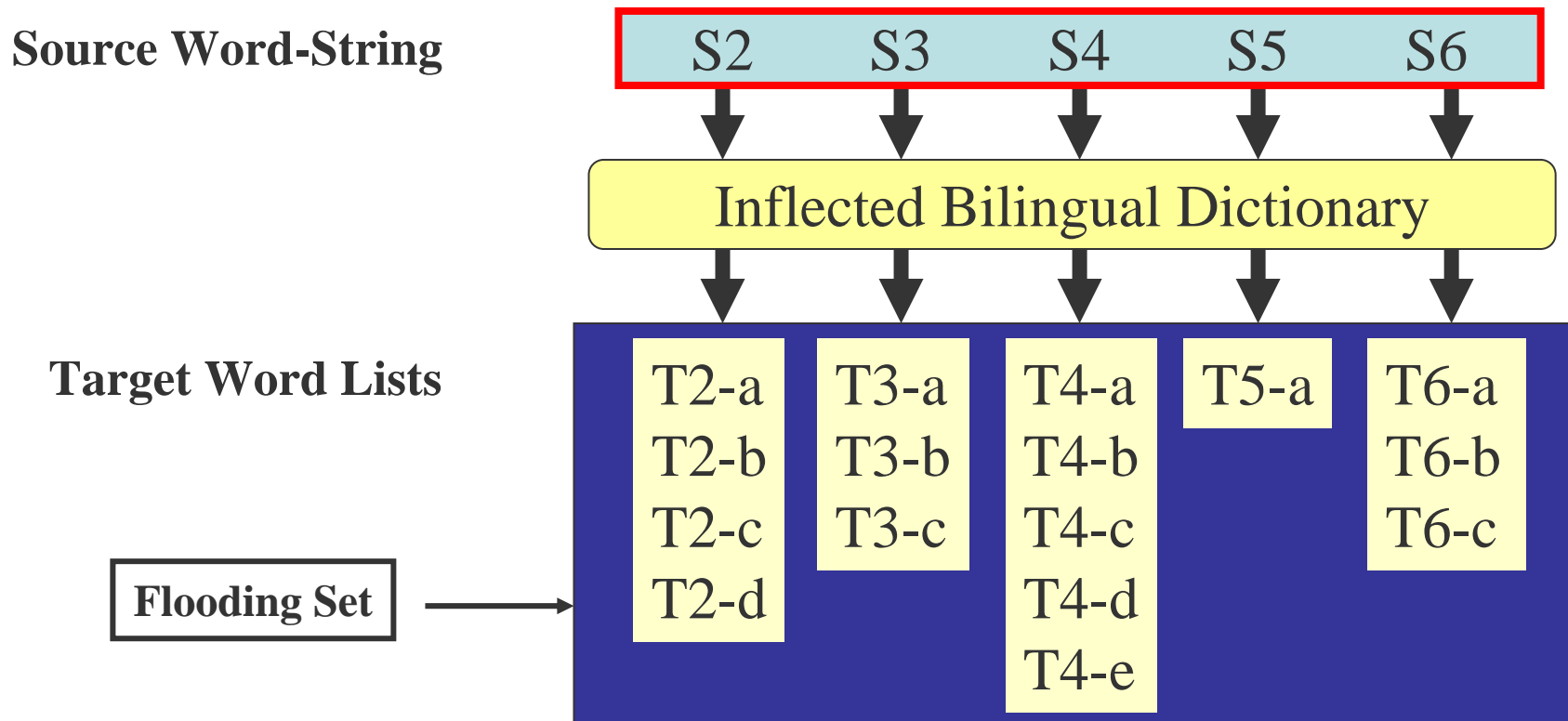
Step 1: Source Sentence Chunking

- 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)



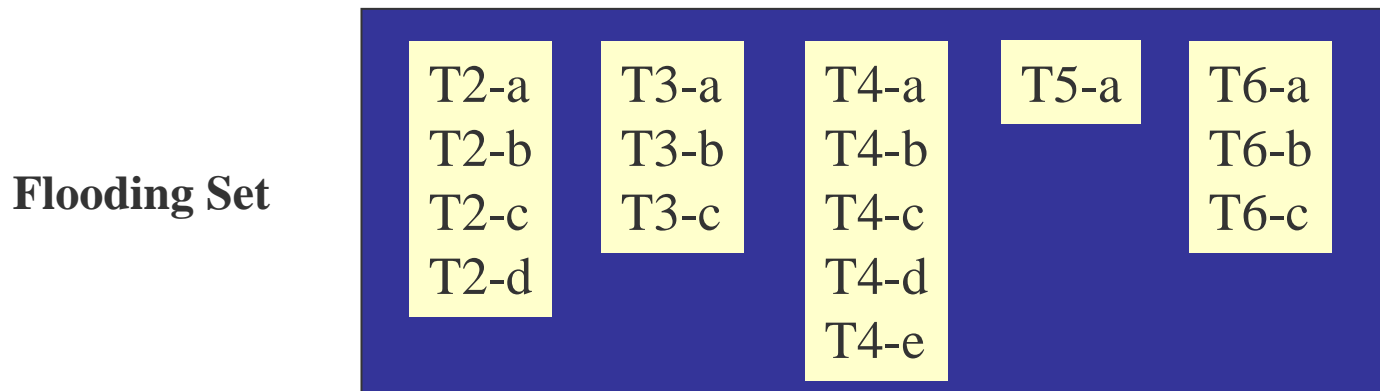
Step 2: Dictionary Lookup

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



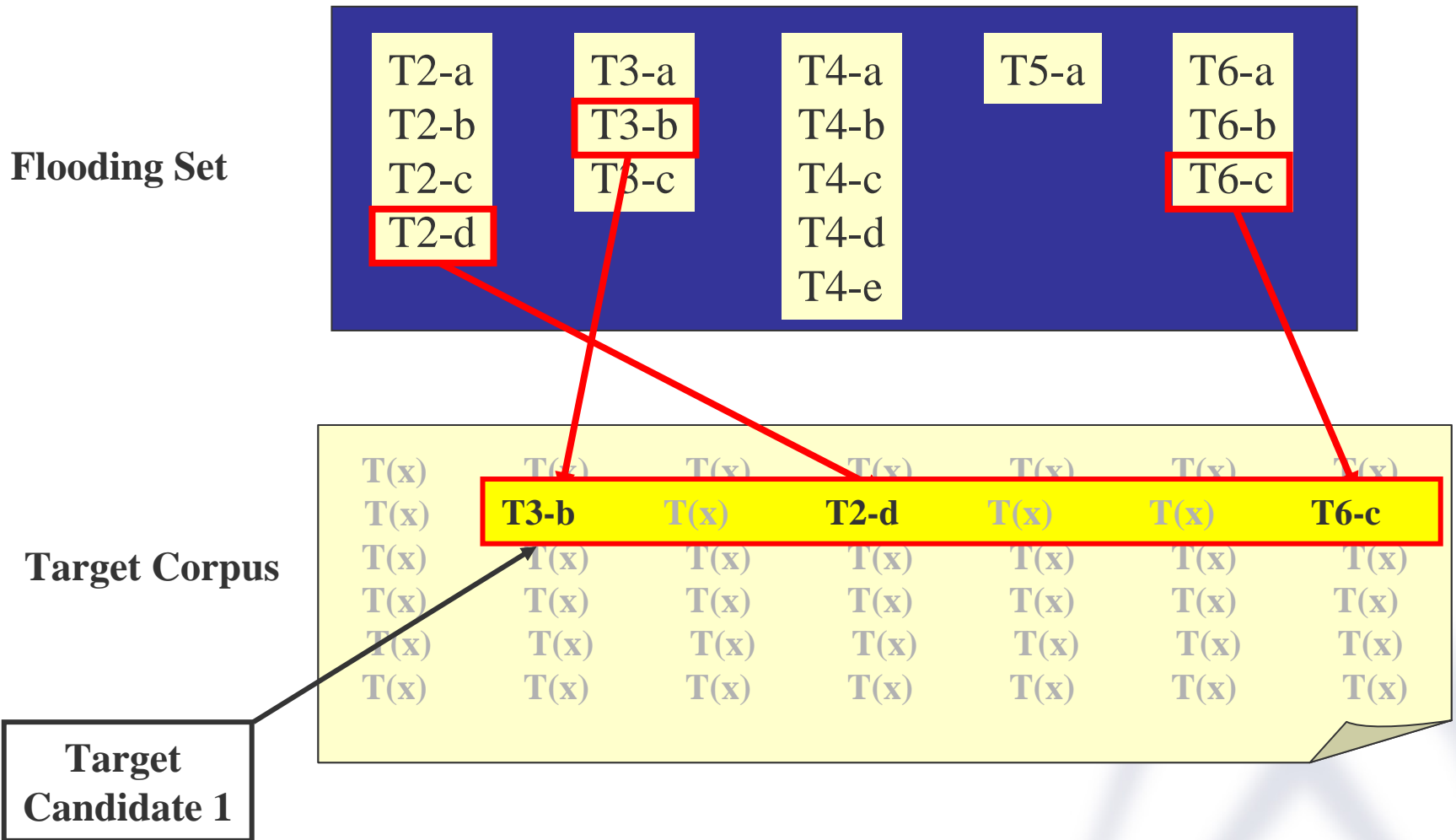
Step 3: Search Target Text

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

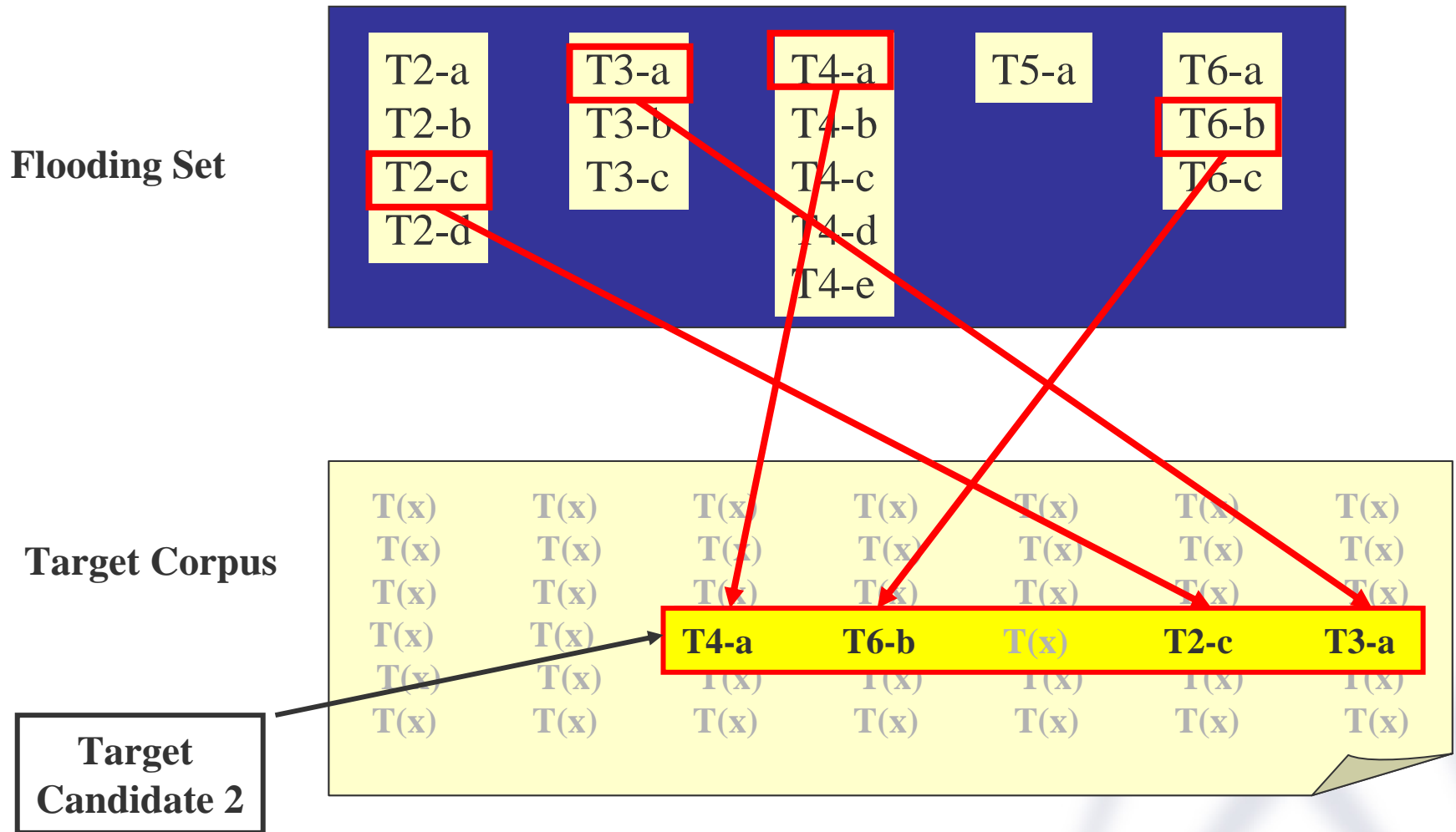


- Find maximum number of words from Flooding Set in minimum length word-string
 - *Words or phrases can be in any order*
 - *Ignore function words in initial step (T5 is a function word in this example)*

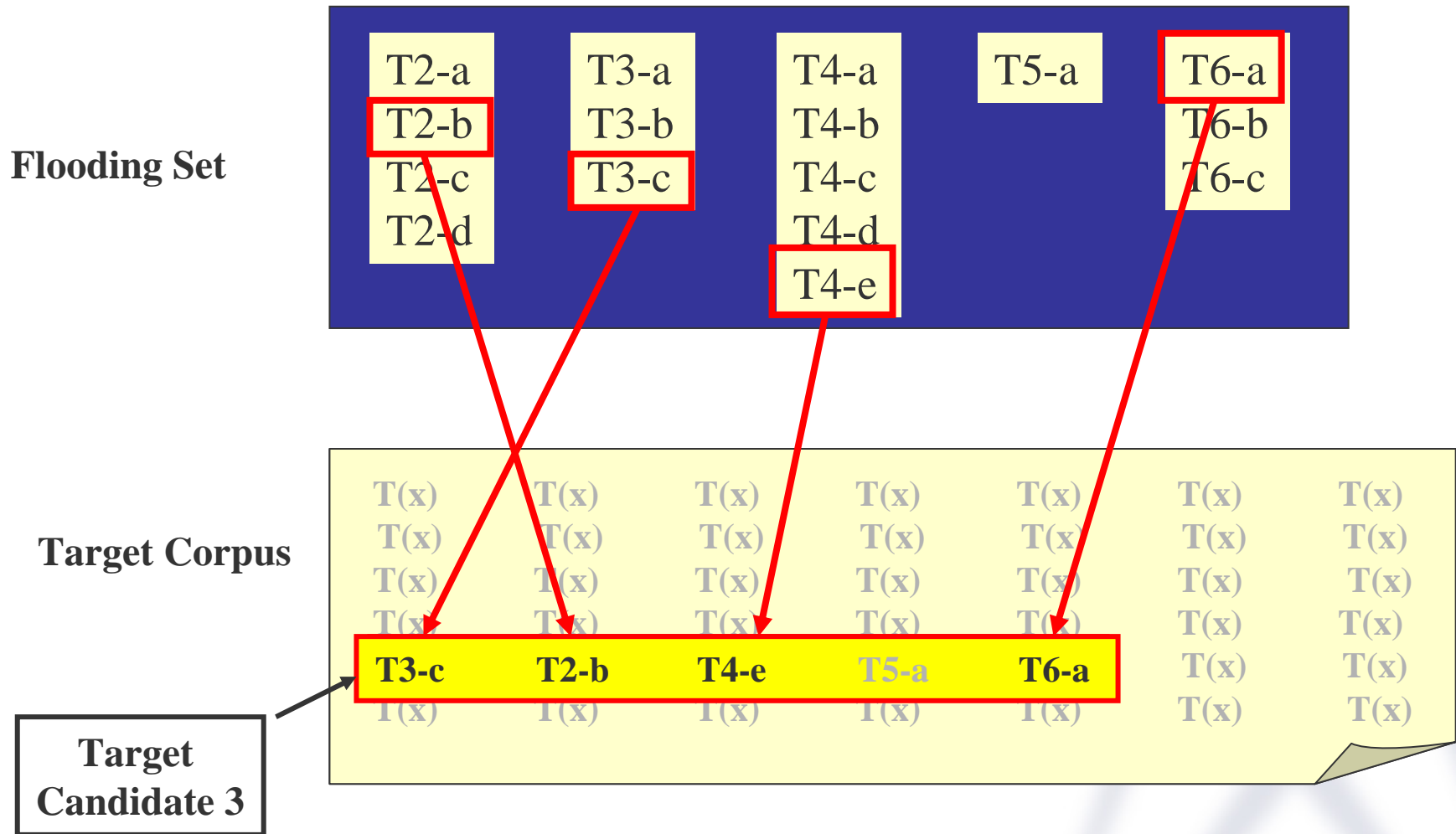
Step 3: Search Target Text (Example)



Step 3: Search Target Text (Example)



Step 3: Search Target Text (Example)



Step 4: Score Word-String Candidates

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

Target Word-String Candidates

T3-b	T(x)	T2-d	T(x)	T(x)	T6-c	3rd
T4-a	T6-b	T(x)	T2-c	T3-a		2nd
T3-c	T2-b	T4-e	T5-a	T6-a		1st

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	

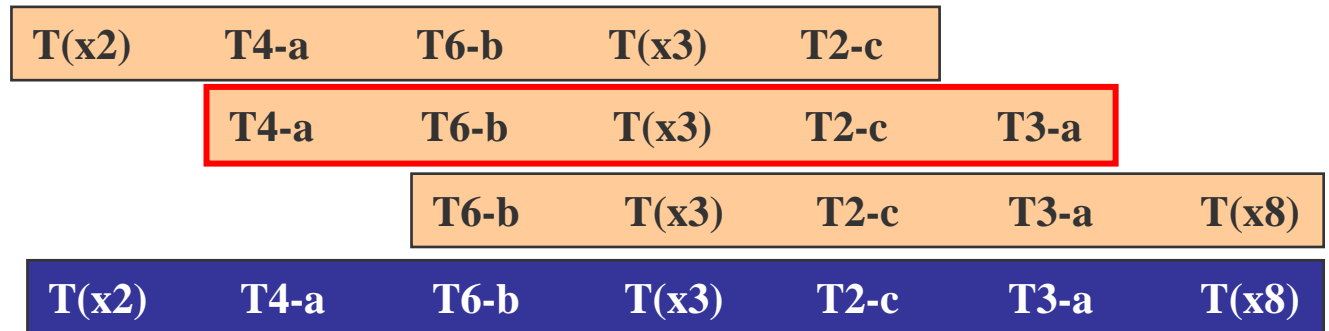
Word-String 3
Candidates

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)

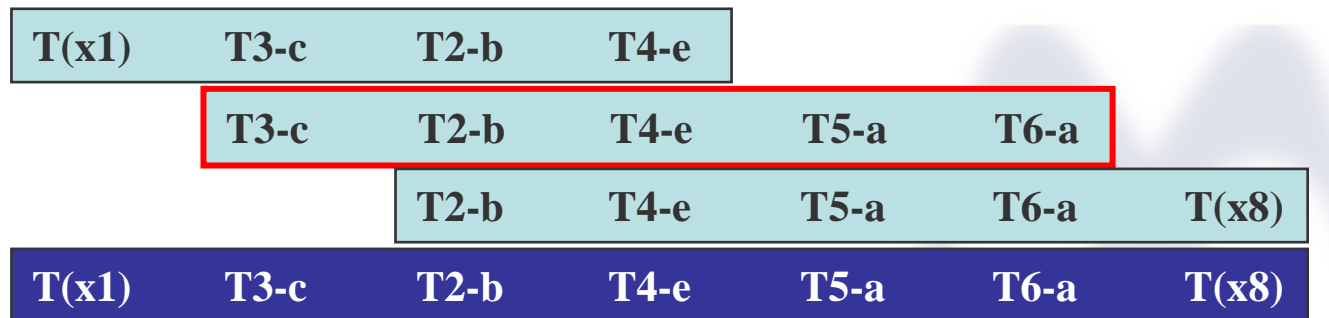
Step 5: Select Candidates Using Overlap

Best translations selected via maximal overlap

Alternative 1



Alternative 2



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

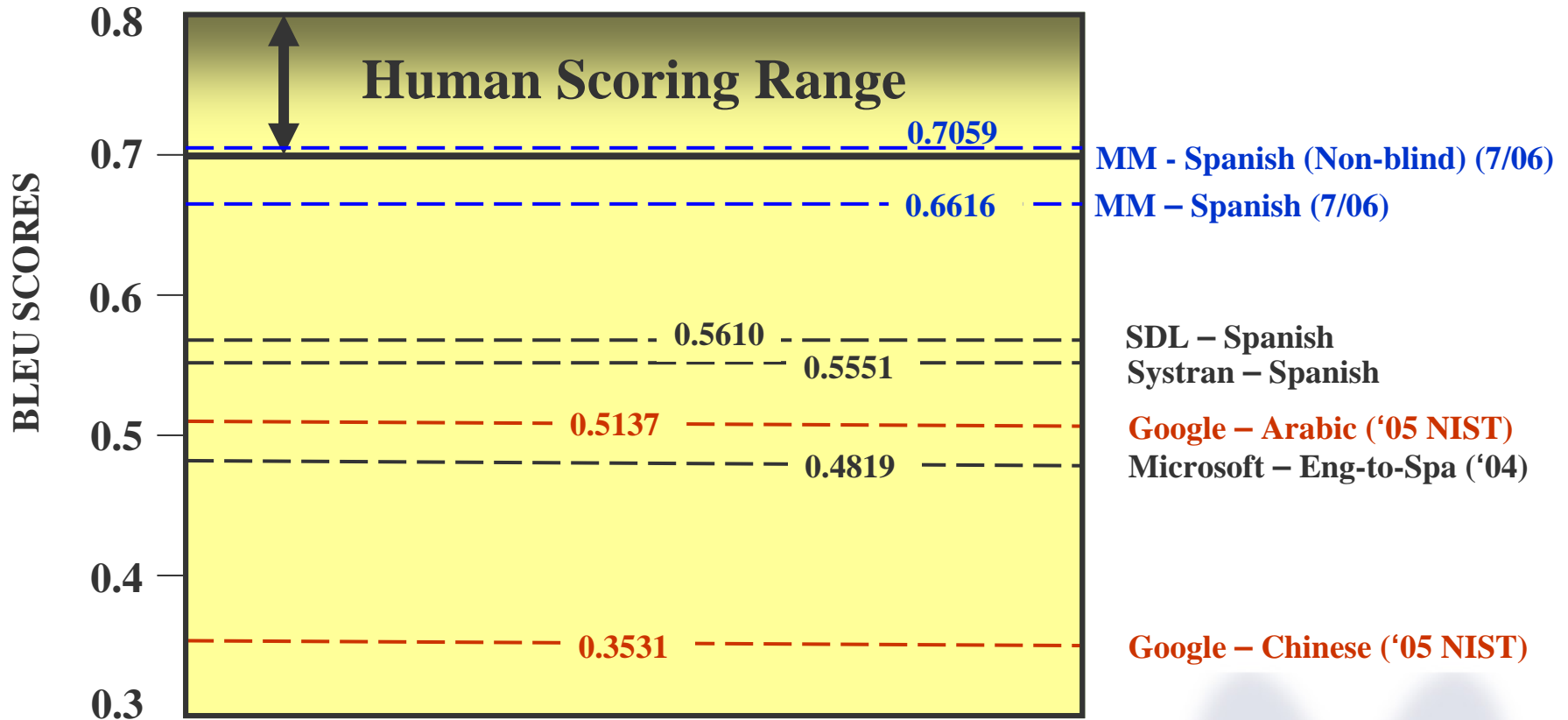
N-grams connected
via Overlap

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

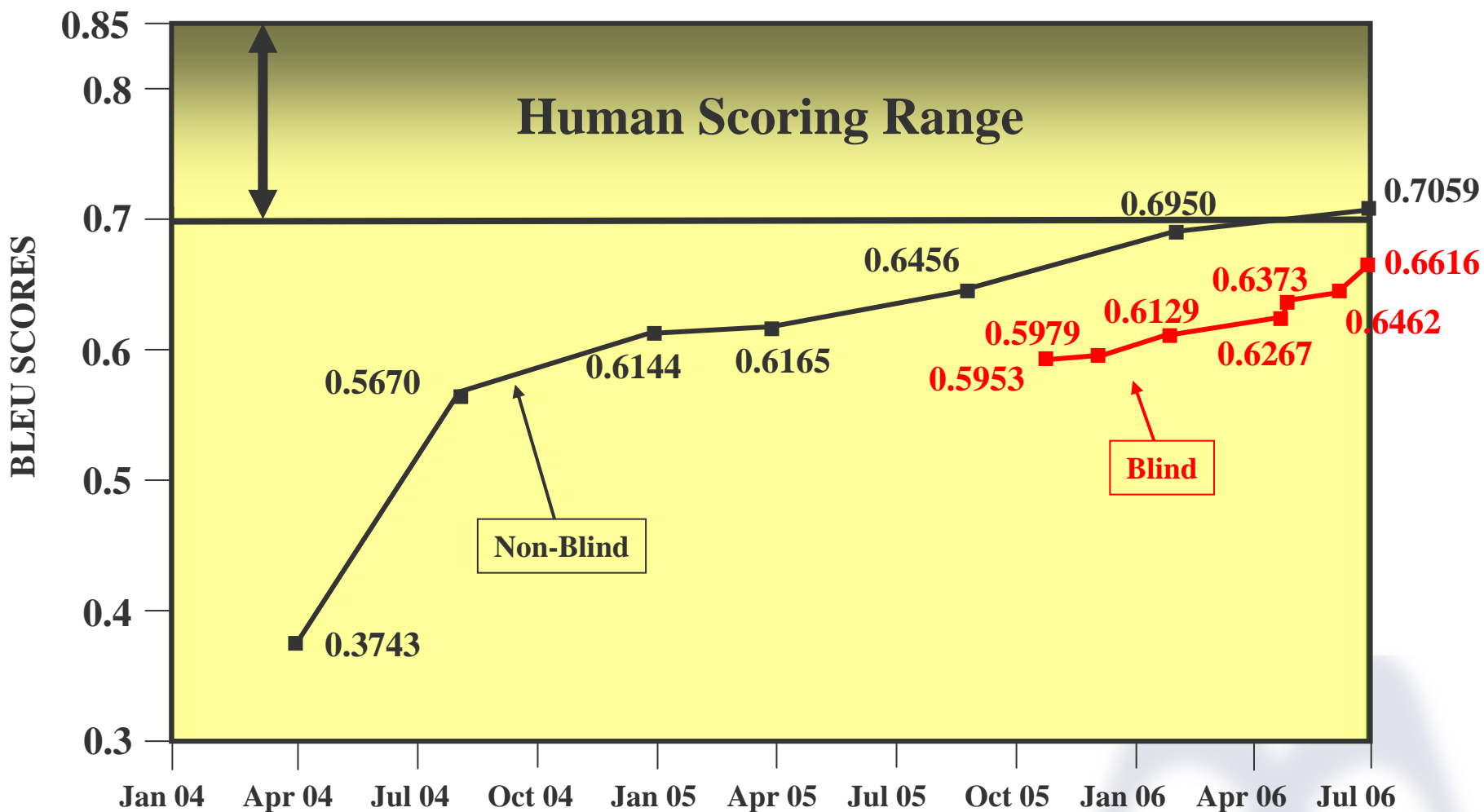
System Scores



“*MM-Spanish*” based on MM’s Spanish-to-English prototype using *incomplete resources*.

“*MM-Spanish (Non-blind)*” based on development testing when expanded language resources are simulated: (1) a larger target corpus and (2) a dictionary that contains all source words (although maybe not all translation senses).

CBMT Scoring Over Time



Note: Recent Blind tests (0.6616, 0.6462) used 53GB of English text. Blind test in April 2006 (0.6373) used 42GB. Previous Blind tests used 30GB. Deltas from bigger corpus & algorithmic improvements.

Illustrative Translation Examples

un coche bomba estalla junto a una comisaría de policía en bagdad

MM: a car **bomb** explodes next to a police station in baghdad

Systran: a car **pump** explodes next to a police station of police in bagdad

hamas anunció este jueves el fin de su cese del fuego con Israel

MM: hamas announced thursday **the end** of the cease fire with israel

Systran: hamas announced **east** Thursday **the aim** of its cease-fire with Israel

testigos dijeron haber visto seis jeeps y dos vehículos blindados

MM: **witnesses** said they have seen **six jeeps** and two armored vehicles

Systran: **six witnesses** said to have seen **jeeps** and two armored vehicles

A More Complex 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
-
- **MM:** 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**

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

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 volatile 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 (94)
stock exchange (92)
national market (89)
national market system (86)
stock market® (85)
exchange (81)
major stock market (61)
the stock market (48)
stock exchange never (32)
stock exchange and (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

Association Builder: Breathing

terrorist organization

1. terrorist network
2. terrorist group
3. terrorist organisation
4. network
5. terror network
6. organization
7. militant group
8. al Qaeda organization

terrorist network

1. terror network
2. terrorist group
3. terrorist organisation
4. network
5. terror group
6. al qaeda network
7. terrorist networks
8. militant group

terrorist group

1. militant group
2. terrorist network
3. group
4. terror group
5. terror network
6. terrorist organization
7. militant network
8. network

terrorist organisation

1. terrorist organization
2. terror network
3. terrorist network
4. terrorist group
5. network of
6. worldwide
7. cadre of

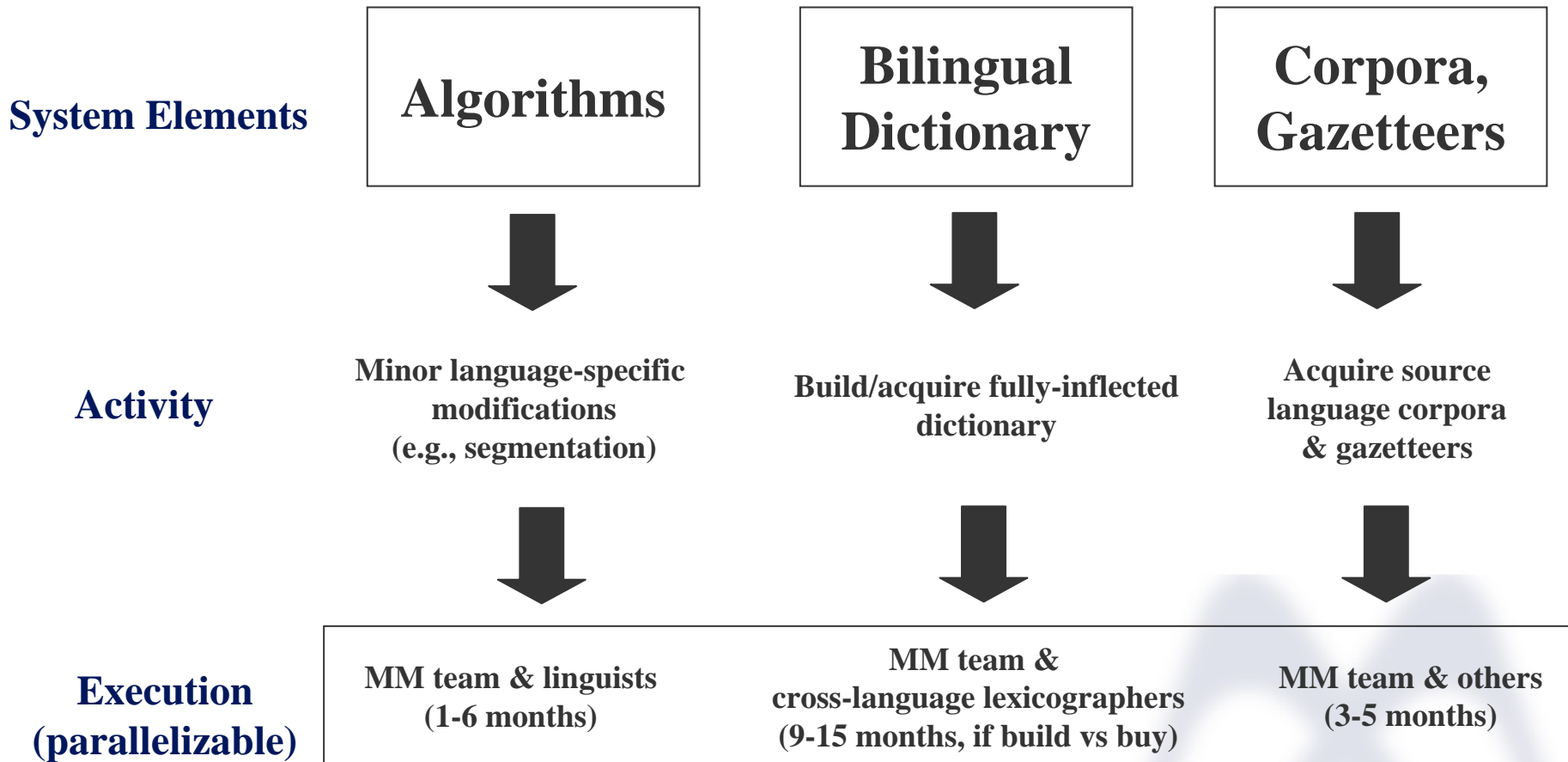
terror network

1. terrorist network
2. network
3. terrorist group
4. terror group
5. terrorist organization
6. militant group
7. militant network
8. terrorist organisation

militant group

1. group
2. terrorist group
3. extremist group
4. militant
5. terror group
6. rebel group
7. resistance movt.
8. Islamist group

Bringing New Languages Online



CBMT Updates since Summer 2006

- Algorithmic improvements since 8/2006 😊
 - Spanish “blind” test BLEU .66 → .69
 - Spanish “non-blind” BLEU .70 → .76
- Spanish-English dictionary 😐
 - Full-form, world’s largest, but needs cleanup
- English (TL) corpus 😐
 - Now: 50GB – 150GB general → Desired: 1TB+,
general + domain segmented
- Run-time 😞
 - 1/2006: 30 sec/word → 1/2007: 3 sec/word →
Desired: 0.01 sec/word (on single server)

Concluding Remarks on CMBT

- **CBMT is truly a new MT paradigm**
 - *Requires no transfer rules and no parallel text*
 - *Achieved very high BLEU scores in S-to-E MT (0.66)*
 - *Well-suited for rare-languages (...but needs bilingual dictionary)*
- **Next Steps for CBMT**
 - CMBT may be what the doctor ordered for distant language pairs such as Chinese-English (still to be determined)
 - Overlap decoder may boost “classical” MT-paradigms: EBMT, SMT, RMT, e.g. starting from MT lattice (still to be determined)
 - Needs more language pairs, larger TL corpus, significant speed-up, robust re-engineering (Chinese-English MT just started)
 - MM willing to let CMU use/modify CBMT source (I think)

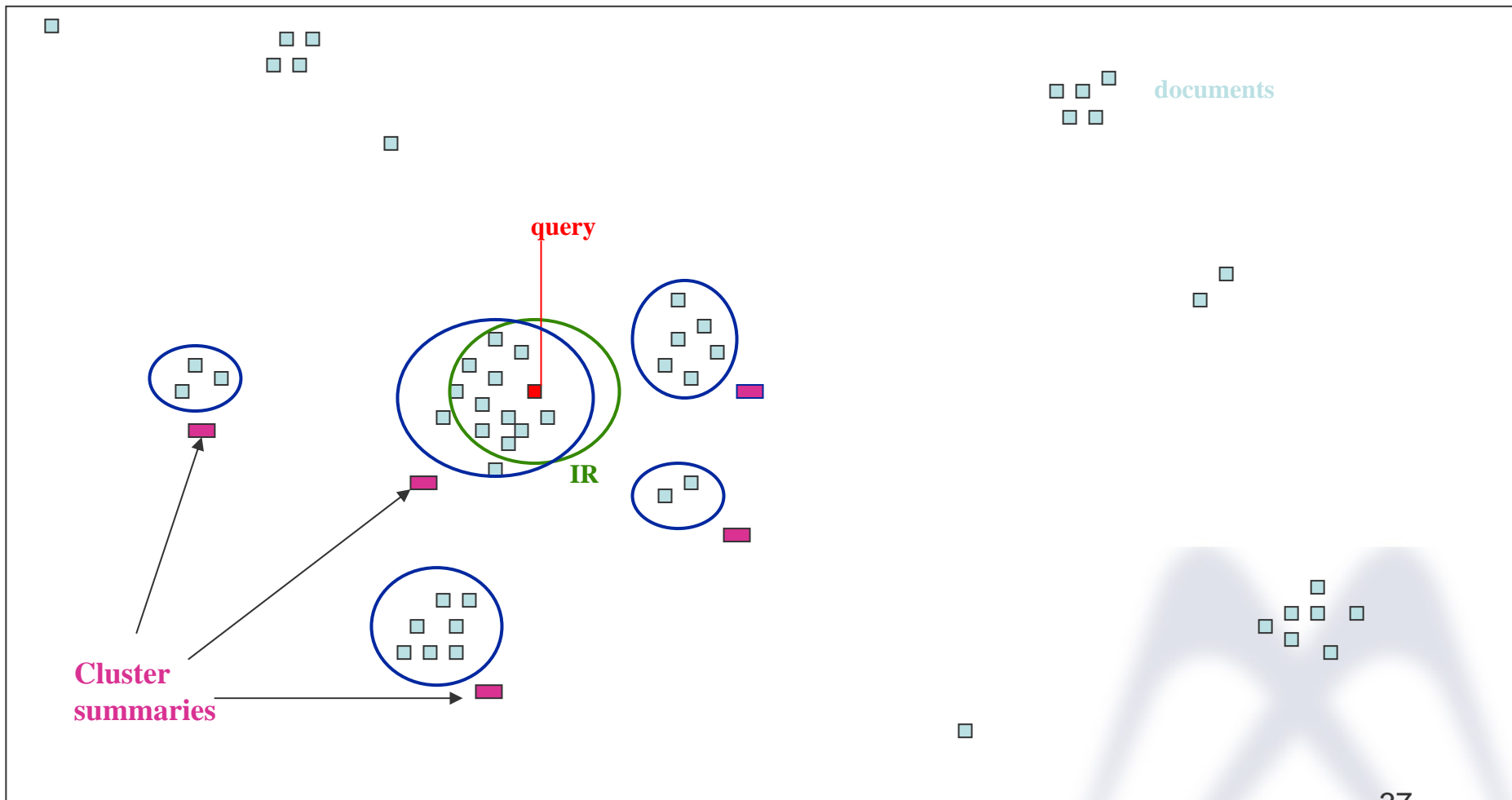
Next Generation Search & Beyond

- Automated Summarization
 - *Multi-document summaries*
 - *User-controllable (length, type, etc.)*
- Document Clustering
 - *Group search results by content similarity*
 - *Then, summarize and label each cluster*
- Personal Profiling
 - *User models (of interests, level of knowledge)*
 - *Task models (progression of types of info needed)*
- Information Push (beyond automated clipping)

NEXT-GENERATION SEARCH ENGINES

- Search Criteria Beyond Query-Relevance
 - **Popularity** of web-page (*link density, clicks, ...*)
 - Information **novelty** (*content differential, recency*)
 - **Trustworthiness** of source
 - **Appropriateness** to user (*difficulty level, ...*)
- “Find What I Mean” Principle
 - Search on semantically related terms
 - Induce user profile from past history, etc.
 - Disambiguate terms (e.g. “Jordan”, or “club”)
 - From generic search to helpful E-Librarians

Clustering Search vs Standard Search (e.g. clusty.com)



NEXT-GENERATION SEARCH: Maximal Marginal Relevance Principle

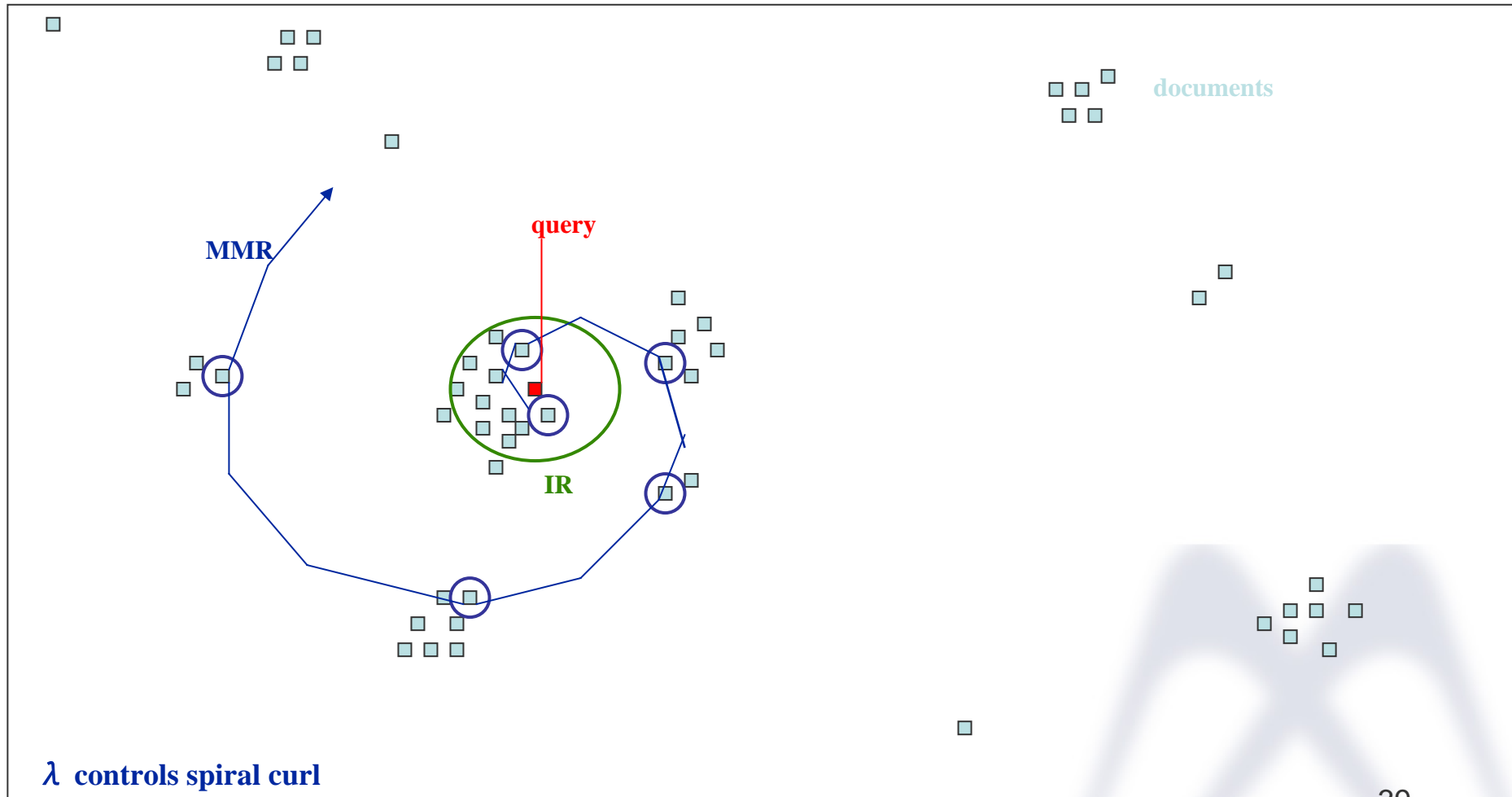
In general, we want to retrieve the k maximally-useful docs,
Where utility = $F(\text{relevance, novelty, popularity, clarity, ...})$

So far, we can do relevance & popularity. Novelty is next,
by defining “marginal relevance” to be “relevant + new”:

$$\text{MMR}(\vec{q}, D, k) = \underset{\vec{d}_i \in D}{\text{Argmax}} [k, \lambda \text{Sim}(\vec{d}_i, \vec{q}) - (1 - \lambda) \max_{\vec{d}_i \neq \vec{d}_j} \text{Sim}(\vec{d}_i, \vec{d}_j)]$$

MMR is used for ranking search results, or for selecting optimal passages in summary generation.

MMR Ranking vs Standard IR



Personalized Context Agents

- Google does not differentiate among users
 - *Elementary school student* Q: “heart cures”
 - *MD Medical researcher* Q: “heart cures”
- User model
 - *Volunteered and learned information*
 - *Constantly updating*
- Beyond the results list for use
 - *Clusters, summaries, synthetic documents*
 - *Customized on the spot, connected to past knowledge*

NEXT-GENERATION SEARCH: Seeking the “Invisible” Web

- **Invisible Web = DB’s Accessible via Web Pages**
 - *Dynamically-generated web-pages from DB’s*
 - *Information (dynamic pages) served via Java apps*
 - *10 to 100 times larger than static HTML web*
 - *Growing faster than static “visible” web*
- **Need Distributed-IR Model to Access (Callan)**
 - *Either unify content or model each DB*
 - *User’s query => appropriate DB(s) => secondary search => unify results*

KNOWLEDGE MAPS:

Search++ for eLibraries

Query: "Tom Sawyer"

RESULTS:

Tom Sawyer home page

The Adventures of Tom Sawyer

Tom Sawyer software (graph search)

Disneyland – Tom Sawyer Island

WHERE TO GET IT:

Universal Library: free online text & images

Bibliomania – free online literature

Amazon.com: The Adventures of Tom...

DERIVATIVE & SECONDARY WORKS:

CliffsNotes: The Adventures of Tom...

Tom Sawyer & Huck Finn comicbook

"Tom Sawyer" filmed in 1980

A literary analysis of Tom Sawyer

RELATED INFORMATION:

Mark Twain: life and works

Wikipedia: "Tom Sawyer"

Literature chat room: Tom Sawyer

On merchandising Huck Finn and Tom Sawyer