Enriching Word Alignment with Linguistic Tags

Linguistic Data Consortium, IBM

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, Kazuaki Maeda

{xuansong, sgrimes, strassel, maeda} @ldc.upenn.edu

niyuge@us.ibm.com

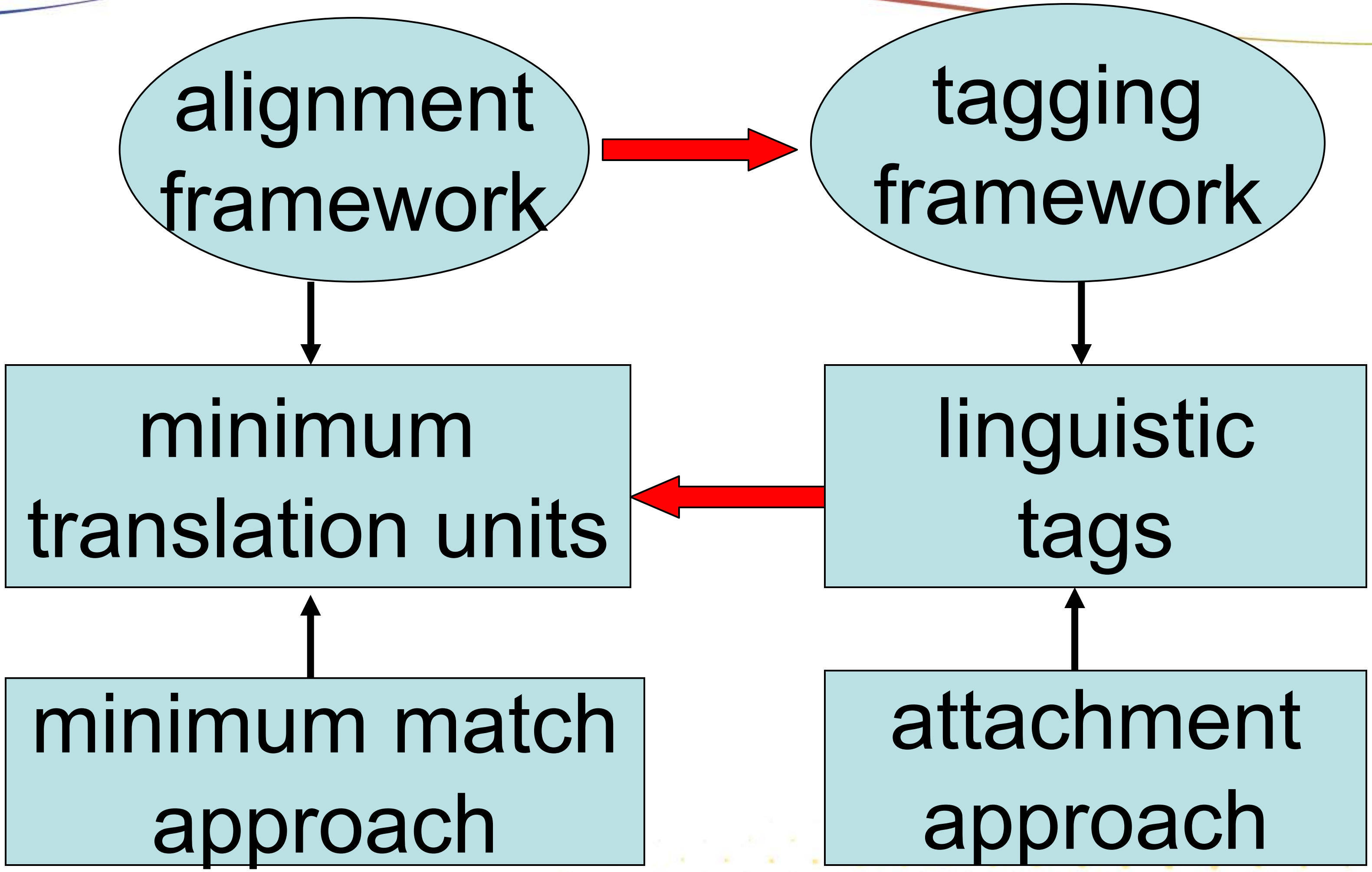
Outline

- Motivations
- Approaches and methodologies
- Linguistic tags
- Inter-annotator agreement
- Conclusions

Motivations

- To improve automatic word alignment quality
- To reduce data amount needed for statistic models
- Supervised models outperform traditional models
- A part of GALE by DARPA: manually aligned and tagged data – Chinese-English WA

Unified Annotation Scheme



alignment framework

tagging framework

minimum translation units

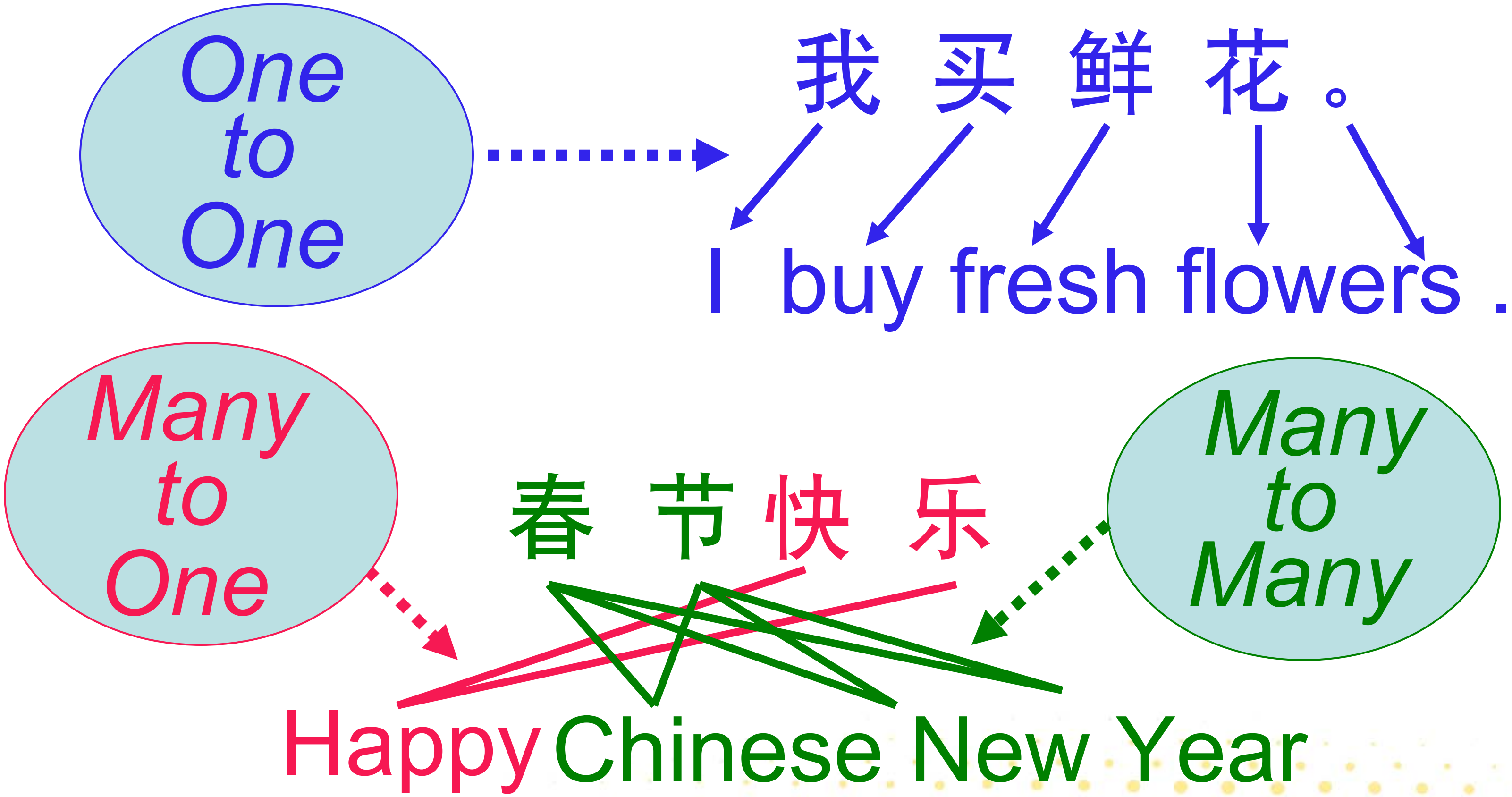
linguistic tags

minimum match approach

attachment approach

Minimum Match Approach

Minimum translation units: atomic



One to One

我 买 鲜 花 。

I buy fresh flowers .

Many to One

Many to Many

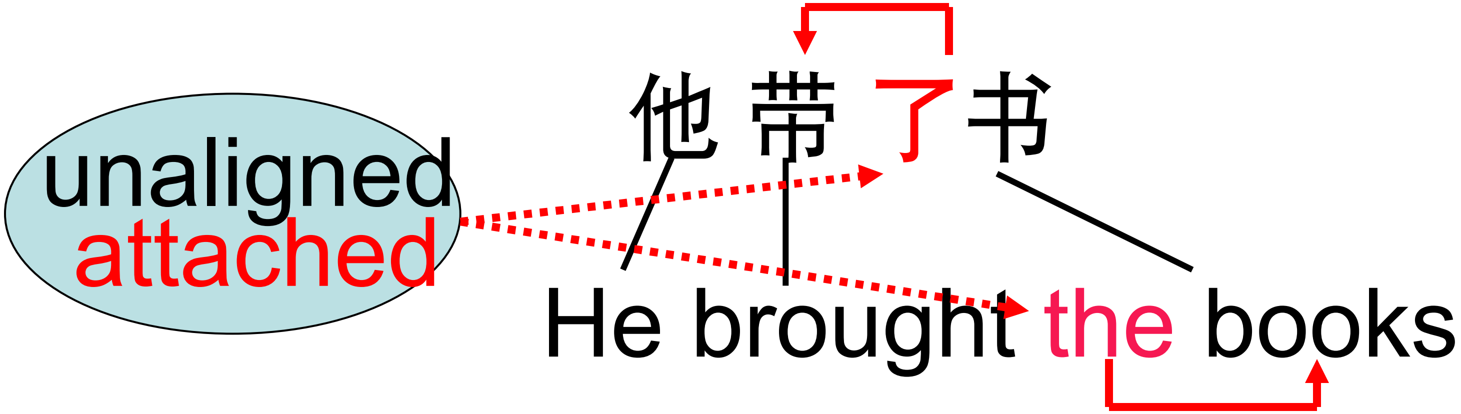
春 节 快 乐

Happy Chinese New Year

Attachment Approach

-- for unaligned words

Attach phrase-level unaligned words

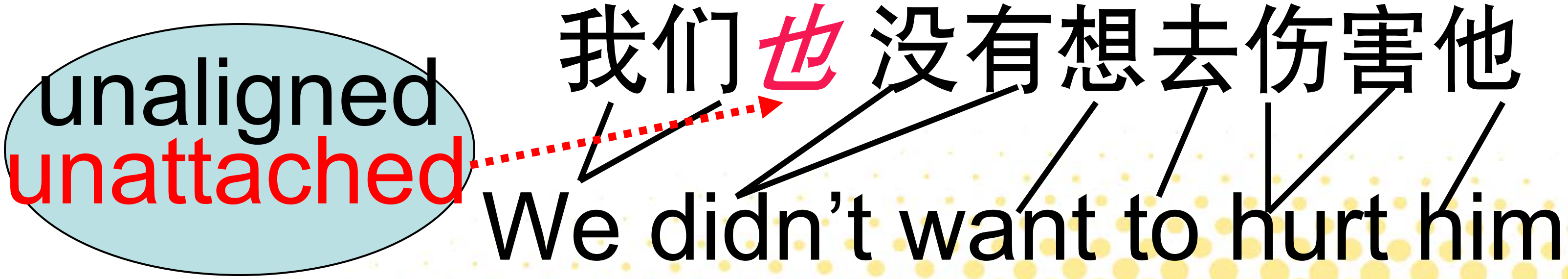


unaligned
attached

他 带 了 书

He brought the books

Unattach sentence-level/discourse-level unaligned words



unaligned
unattached

我们也没有想去伤害他

We didn’t want to hurt him

Tagging Framework

Goal: tackle insertion/deletion problems

Methodologies: using linguistic tags

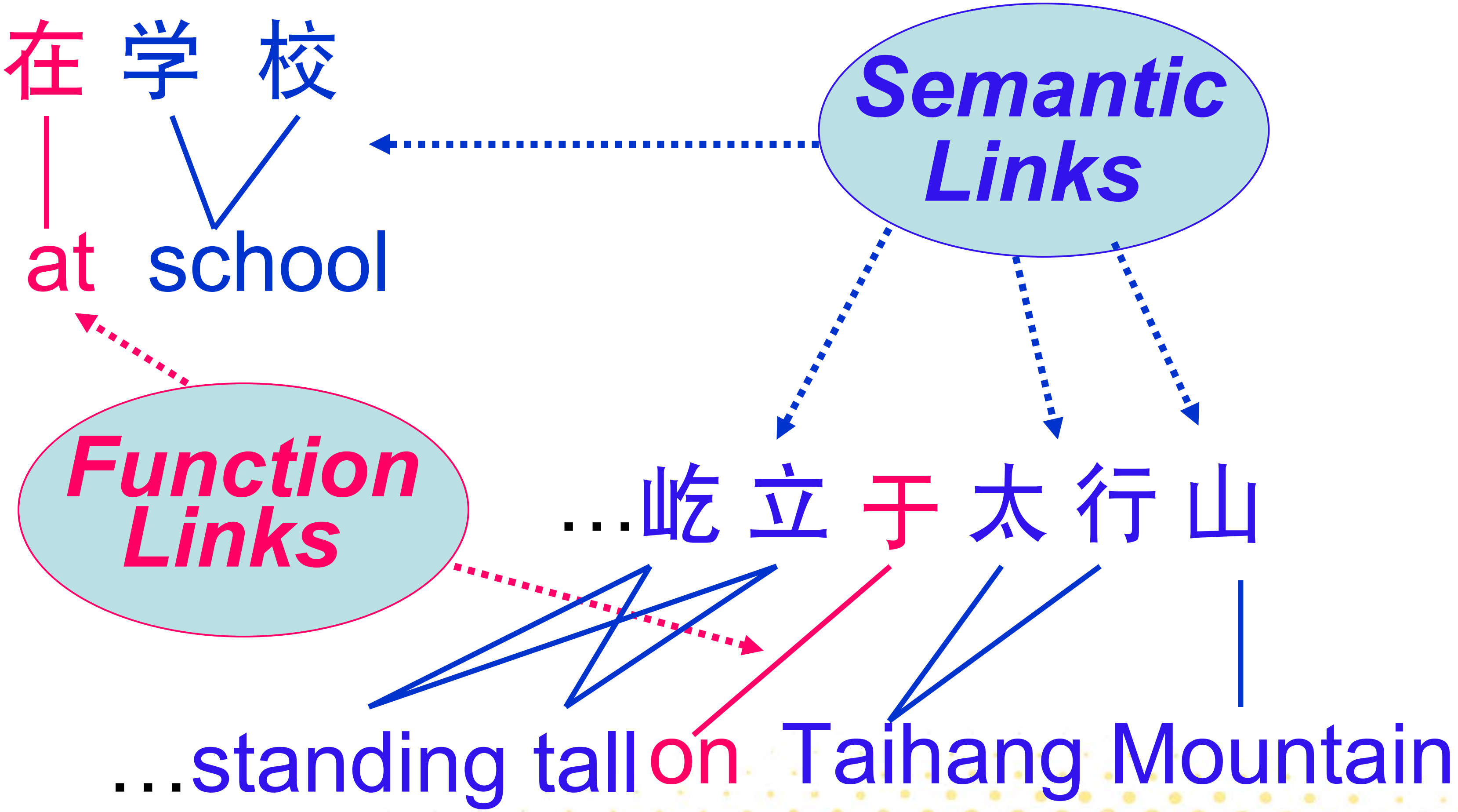
-Tag aligned links

- Context-free links (2)
- Context-dependent links (3)
- Specific-feature links: Chinese-DE的 (3)

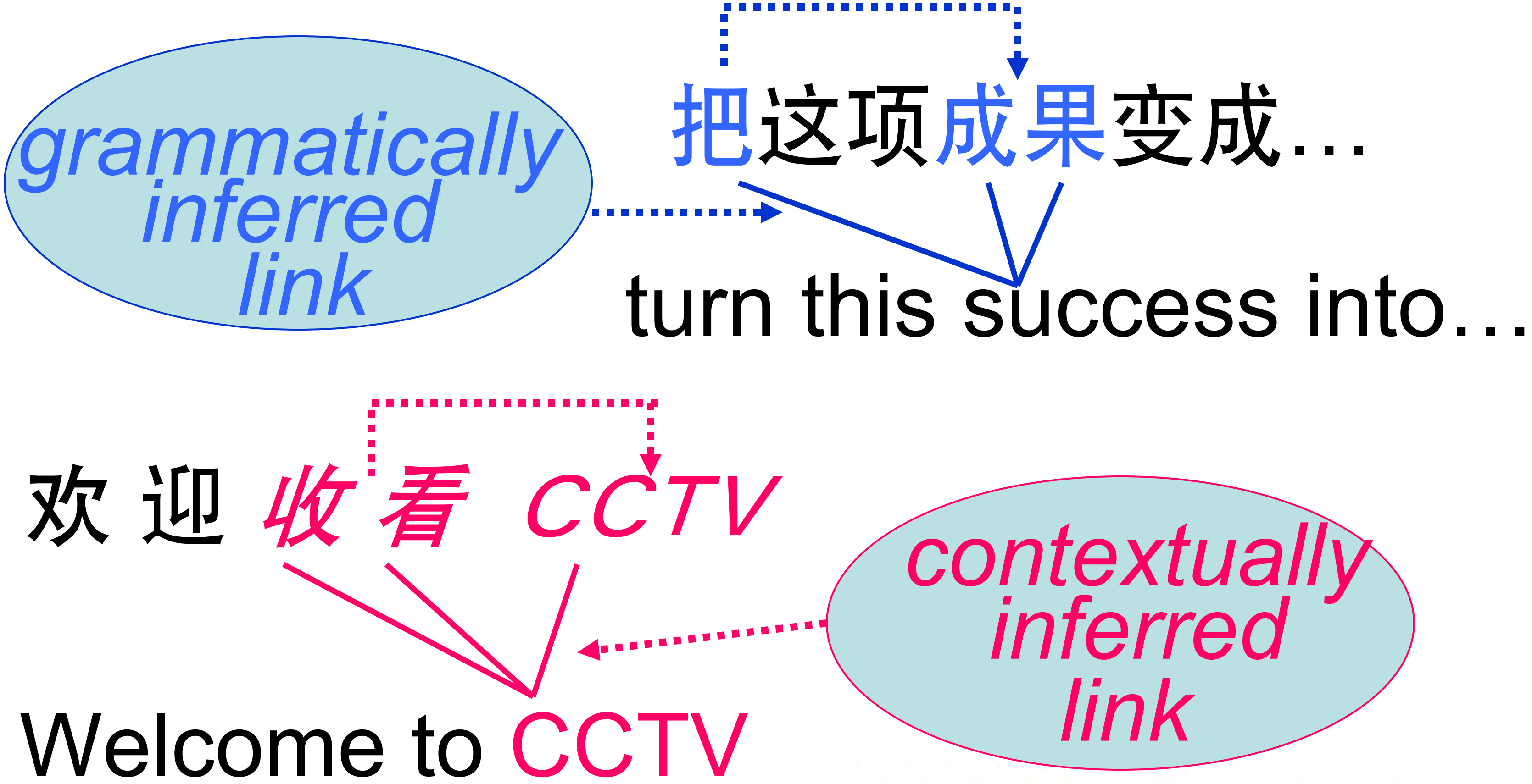
-Tag unaligned words

- Tags for attached words(12 types)
- Tags for unattached words (2 types)

Context-free Links



Context-dependent Links



grammatically inferred link

把这项成果变成…

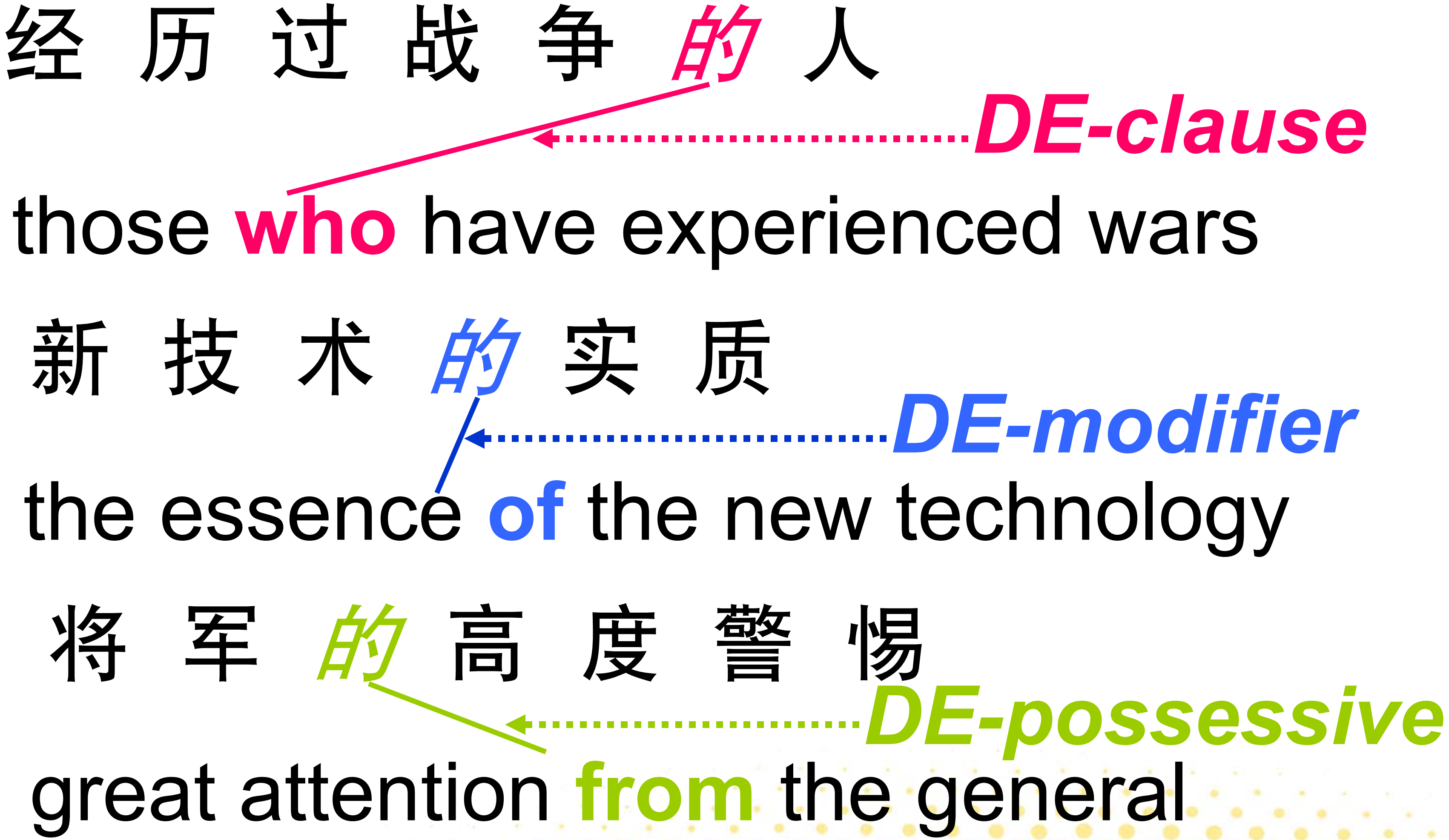
turn this success into…

欢 迎 收 看 CCTV

Welcome to CCTV

contextually inferred link

Specific Links: 的(DE)



经 历 过 战 争 的 人

DE-clause

those **who** have experienced wars

新 技 术 的 实 质

DE-modifier

the essence **of** the new technology

将 军 的 高 度 警 惕

DE-possessive

great attention **from** the general

Word Tags

Aligned & Unaligned

Omni-func-preposition	Tense/Passive
Possessive	Measure word
Clause marker	Rhetorical
Sentence marker	Co-reference
Determiner	TO-infinitive
DE-modifier	Local context
Context-obligatory	
Non-context-obligatory	

Examples: Word Tags

Word Tag	Examples
Possessive	the head **of** the branch
Measure-word	一根(one) 柱子 (pillar) [one pillar]
Tense/Passive	提交(submit)的报告(report) [report submitted]
Context-obligatory	不(not)好(easy)掌握(control),凭(by)经验(experience) [**It** is not easy to control, **you do** by experience]
Non-context-obligatory	他(he)都已经(already)签(sign) 合同了(contract) [He already signed a contract]

Inter-Annotator Agreement(1)

Chinese-English Alignment

Data Source	*Char-Count*	*Precision*	*Recall*	*F-score*
NW1	306	97.3%	95.7%	96.5%
NW2	185	95.3%	96.2%	95.7%
NW3	365	90.4%	91.2%	90.8%
NW4	431	90.8%	92.6%	91.2%

Inter-Annotator Agreement(2)

Chinese-English Tagging

Data Source	*Chi. Char*	*Eng. Word*	*Link Count*	*Same Tag*	*Agree*
NW1	306	233	186	683	94.2%
NW2	185	131	105	392	93.1%

Conclusion

- Unified annotation scheme
- Manually aligned and tagged corpora at LDC
- Annotation guidelines available at:

  http://projects.ldc.upenn.edu/gale/task_specifications/
- Annotation toolkit available soon
- On-going project: more data in pipeline
- Acknowledgements to GALE of DARPA

Thank You!

Chinese-English Aligned and Tagged Corpora at LDC

Genre	*File*	*Char*	*Segment*
Newswire	579	225645	5015
Broadcast News	28	183400	6376
Broadcast Conversation	34	306497	12050
Weblog	747	229799	9382
Total	1388	945341	32823

Linguistic Data Consortium

Annotation Rate

Average skill, speed and difficulty level

- First pass alignment: 10,000w/10h
- Second pass alignment: 10,000w/6h
- First pass tagging: 10,000w/7h
- Second pass tagging: 10,000w/5h