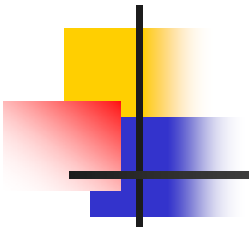
Extracting Equivalent Chunks from Chinese-English Bilingual Corpus

Yanjun Ma

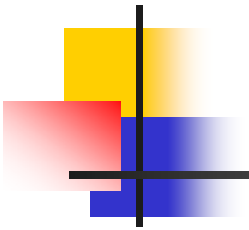
National Centre for Language Technology, Dublin City University

Note: This work is part of my master thesis, Tsinghua University, Beijing, China



Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work



Related research

- Different levels of alignment
  - Passage, paragraph alignment
  - Sentence alignment
  - Word alignment
  - Chunk (phrase, sub-sentential) alignment

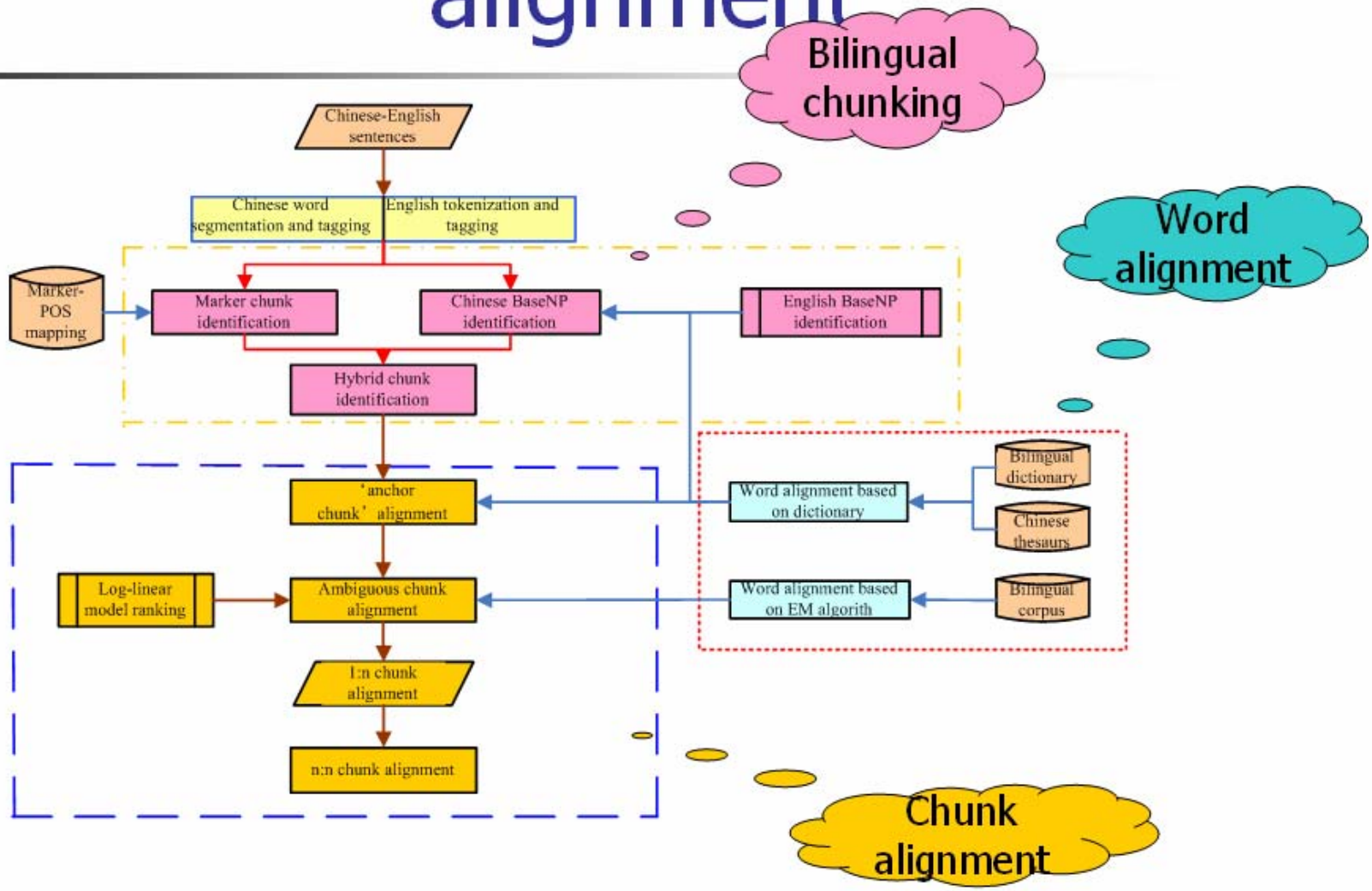
Related Research (cont.)

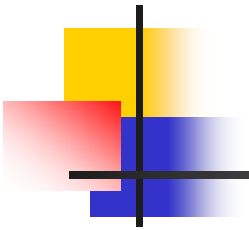
- Generalized chunk styles
  - Consecutive word sequence (SMT)
  - CoNLL-2000 style chunks
  - Marker chunks (EBMT style chunks)
  - Sub trees in a grammar tree

Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work

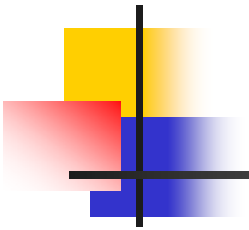
Framework for chunk alignment





Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work



Word alignment---overview

- Statistical V.S Knowledge base
- Statistical approach
  - Heuristic word alignment
  - IBM model 1~5
- Word alignment using knowledge base
  - Bilingual dictionary
  - thesaurus

Word alignment---1:1 alignment algorithm

C={<c_1, 1>,<c_2, 2>, …<$c_{J'}$, J>}; // set of Chinese words
E={<e_1, 1>,<e_2, 2>, …<eI', I>}; // set of English words
A=Φ; // set of word alignment
foreach <$c_{j'}$, j>$\in$C, 1$\leqslant$j'$\leqslant$j$\leqslant$J, <$e_{i'}$, i>$\in$E, 1$\leqslant$i'$\leqslant$i$\leqslant$I{
    if (CEAlignScore ($c_{j'}$, $e_{i'}$) > h)
        add (A,<$c_{j'}$, j, $e_{i'}$, i>); // add an alignment
}
C=C-{<c_j,j>|Exist <$e_{i'}$, i>,<$c_{j'}$, j, $e_{i'}$, i>$\in$A};
E=E-{<ei', i>|Exist <cj', j>, <$c_{j'}$, j, $e_{i'}$, i>$\in$A};
foreach <$c_{j'}$, j>$\in$C{
    S=GetSynonym ($c_{j'}$); // Get synonym given a word
    foreach c_j $\in$S, <$e_{i'}$, i>$\in$E
        if CEAlignScore (c_j, $e_{i'}$) > h
            add (A, <c_j, j, $e_{i'}$, i>);
}
Output A;

Word alignment---1:1 alignment algorithm (cont.)

$$CEAlignScore(c,e) = \begin{cases} 1 & c = e \\ \max\limits_{x \in Dict(c)} EEsim(x,e) & c \neq e \end{cases}$$

$$EESim(x,y) = \begin{cases} 1 & if\ x = y \\ 0.9 & if\ x = WordStem(y) \\ 0.8 & if\ x = SubString(y)\ or\ y = SubString(x) \\ 0 & else \end{cases}$$

- Function description:
  - Dict: Get a set of English translation given a Chinese word
  - EESim: Get the similarity between two English words

Word alignment---word anchors

- Define word anchors (WA) as follows:

$WA = \{\langle c_{j'}, j, e_{i'}, i\rangle | Count(c_{j'}, j, A) = 1$ and $Count(e_{i'}, i, A) = 1\}$

- Ambiguous word alignment (AA)

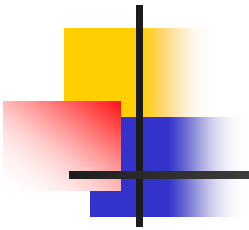
$AA = A - WA$

Word alignment---an example

- 他们$_1$ 中间$_2$ 有$_3$ 几个$_4$ 人$_5$ 懂$_6$ 汉语$_7$。$_8$
  A$_1$ few$_2$ of$_3$ them$_4$ know$_5$ Chinese$_6$.$_7$
- Result of word alignment
  <他们, 1, them, 4>: 0.9
  <汉语, 7, Chinese, 6>: 1.0
- Entries in Chinese-English dictionary
  懂：understand
  知道：know

Word alignment---an example

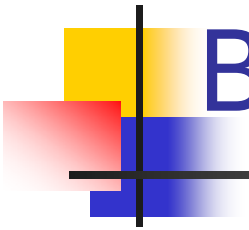
- Word expansion based on Chinese Thesaurus
  - S = GetSynonym (懂) = {知道, 了解……}
  - 知道: know
- Get word alignment:<懂, 6, know, 5>



Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work

Bilingual Chunking---overview



- Marker hypothesis
- Marker chunk identification
- Difficulties in Chinese-English marker chunk alignment
- Hybrid chunks: a combination of marker chunks and BaseNP
- BaseNP identification using bilingual corpus

Bilingual Chunking---Marker Hypothesis

- Green (1979)
- Natural languages are ‘marked’ for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes.

Bilingual Chunking---Marker Hypothesis

Marker sets	examples
<DET>	{the, a, an, those, these, ...} determiners
<PREP>	{in, on, out, with, from, to, under, ...} prepositions
<QUANT>	{all, some, few, many, ...} quantifiers
<CONJ>	{and, or, ...} conjunctions
<POSS>	{my, your, our, ...} possessive pronouns
<PRON>	{I, you, he, she, it, ...} pronouns
.....	……

Marker sets	examples
<DET>	{这，那，该，各，这些，任何……}
<PREP>	{在，由，向，跟，按照，关于……}
<QUANT>	{很多，大量，一些……}
<CONJ>	{不止，除非，要是，……}
<PRON>	{我，你，他，她们……}

Bilingual chunking ---marker chunk identification

- English marker words can be identified without parsing
- Unfortunately, Chinese marker words can't!
- Chinese marker words are extremely ambiguous
- One-one mapping between marker sets and POS tags

  marker word identification ←→ POS tagging

Bilingual chunking ---marker chunk identification

- Mapping between POS tags and marker sets

English POS	Marker set
IN	PREP
CD, OD	QUANT
CC	CONJ
DT	DET
PRP	PRON
PRP$	POSS

Chinese POS	Marker set
p	PREP
m	QUANT
c	CONJ
r（determinative pronouns）	DET
r（person）	PRON

Bilingual chunking ---marker chunk identification

- An example for marker chunk

抱/v 着/u 两/m 个/q 婴儿/n 的/u 妇女/n 正/d 向/p 托儿所/n 走/v 来/f 。/w
With two baby woman to nursery walk
A/DT woman/NN with/IN two/CD babies/NNS is/VBZ coming/VBG to/TO the/DT nursery/NN ./.

<BEGIN>抱/v 着/u <QUANT>两/m 个/q 婴儿/n 的/u 妇女/n 正/d <PREP>向/p 托儿所/n 走/v 来/f <PUNC>。/w
<DET>A/DT woman/NN <PREP>with/IN two/CD babies/NNS is/VBZ coming/VBG to/TO <DET>the/DT nursery/NN <PUNC>./.

Bilingual chunking---difficulties in Chinese-English chunk alignment

- Troubles in Chinese-English marker chunk alignment

Marker chunks

<BEGIN>抱/v 着/u <QUANT>两/m 个/q 婴儿/n 的/u 妇女/n 正/d <PREP>向/p 托儿所/n 走/v 来/f <PUNC>。/w
<DET>A/DT woman/NN <PREP>with/IN two/CD babies/NNS is/VBZ coming/VBG to/TO <DET>the/DT nursery/NN <PUNC>./.

Hybrid chunks

<BEGIN>抱/v 着/u <BASENP>两/m 个/q 婴儿/n <BACKNP>的/u <BASENP>妇女/n <BACKNP>正/d <PREP>向/p <BASENP>托儿所/n <BACKNP>走/v 来/f <PUNC>。/w <BASENP>A/DT woman/NN <PREP>with/IN <BASENP>two/CD babies/NNS <BACKNP>is/VBZ coming/VBG to/TO <BASENP>the/DT nursery/NN <PUNC>./.

<BEGIN>抱/v 着/u: <PREP>with/IN
<BASENP>两/m 个/q 婴儿/n: <BASNP>two/CD babies/NNS
<BASENP>妇女/n: <BASENP>A/DT woman/NN
<BASENP>托儿所/n: <BASENP>the/DT nursery/NN

Bilingual chunking ---Hybrid chunks

- Identification of hybrid chunks
  - Marker chunks identification

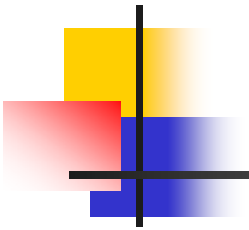
    <DET>A/DT woman/NN <PREP>with/IN two/CD babies/NNS is/VBZ coming/VBG to/TO <DET>the/DT nursery/NN <PUNC>./.
  - BaseNP identification

    <DET> [A/DT woman/NN] <PREP>with/IN [two/CD babies/NNS] is/VBZ coming/VBG to/TO <DET> [the/DT nursery/NN] <PUNC>./.
  - Combination with some priority order

    <BASENP>A/DT woman/NN <PREP>with/IN <BASENP>two/CD babies/NNS <BACKNP>is/VBZ coming/VBG to/TO <BASENP>the/DT nursery/NN <PUNC>./.

Bilingual chunking---BaseNP identification using bilingual corpus

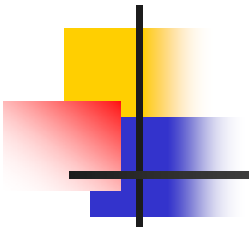
- BaseNP identification using bilingual corpus
  - English BaseNP identification
  - Generation of Chinese BaseNP candidates
  - Chinese BaseNP selection



Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work

Log-linear model for chunk alignment



- Two-step hybrid chunk alignment
  - Unambiguous chunks (Anchor chunks) --- heuristics
  - Ambiguous chunks --- ranking with a log-linear model

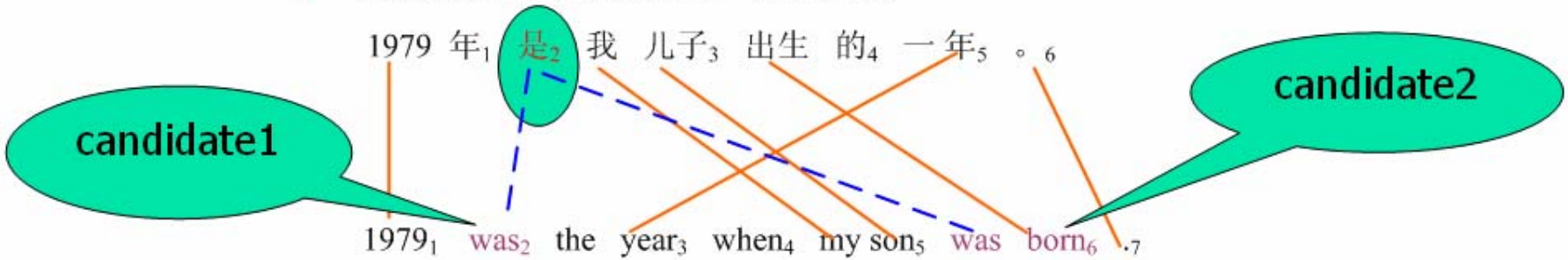
Log-linear model for chunk alignment (cont.)

- Unambiguous chunks (anchor chunks)
  - Premise: anchor words between Chinese chunk and English chunk
  - Constraint: translation of each Chinese words in Chinese chunk is in corresponding English chunk, vice versa
- Ambiguous chunks

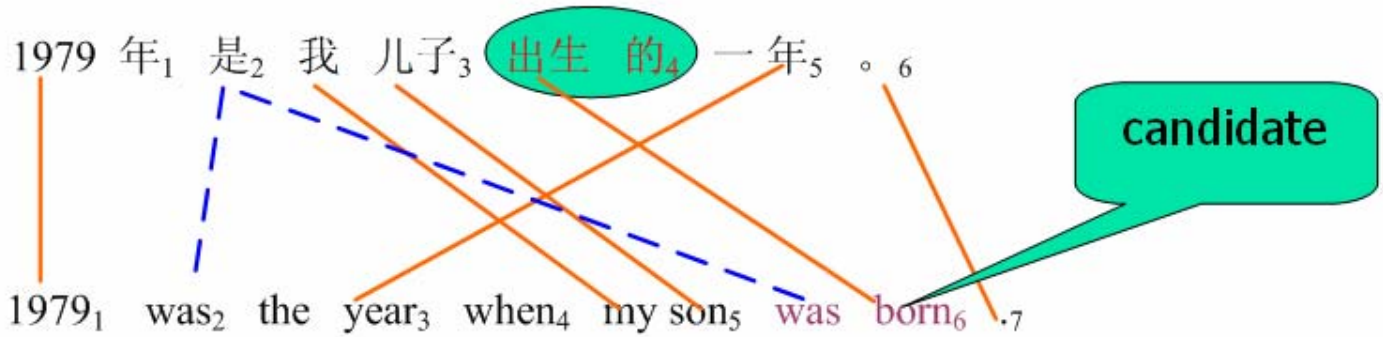
Log-linear model for chunk alignment (cont.)

Why?

- Ranking with a log-linear model
  - Without anchor words



  - Constraints detection



Log-linear model for chunk alignment (cont.)

- Ranking based on log-linear model
  - Candidate generation
  - Feature selection
    - Scores of word alignment
    - Distance distortion
    - Marker transition probability
  - Feature weighting

Log-linear model for chunk alignment (cont.)

- Feature selection and parameter estimation
  - Score of word alignment（F-WALI）

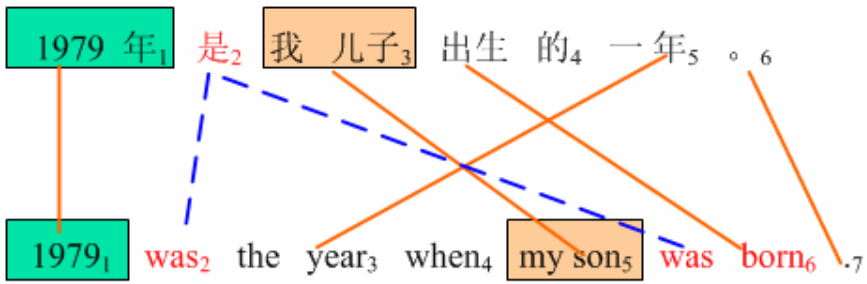
$$Set_W = \{< c_{j'}, j', 1, e_{i'}, i', 1 > | \; lwc \le j' \le rwc, \; lwe \le i' \le rwe, \; < c_{j'}, j', 1, e_{i'}, i', 1 > \in Set_A\}$$

$$Set_{W'} = \{< c_{j'}, j', 1, e_{i'}, i', 1 > | \; lwc \le j' \le rwc, \; lwe \le i' \le rwe, \; < c_{j'}, j', 1, e_{i'}, i', 1 > \in Set_{EM}\}$$

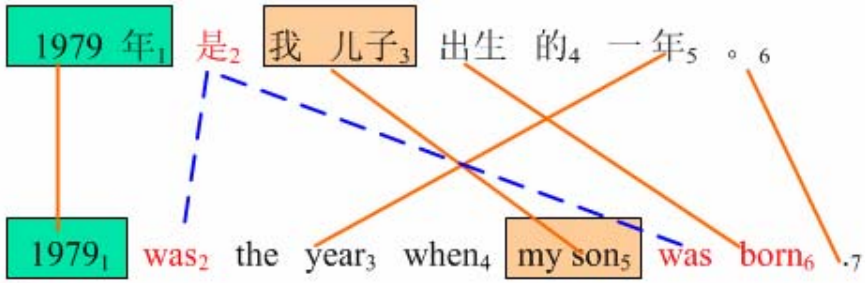
$$F-WALI = \sum_{a \in Set_W} \text{GetProb}(a) + \sum_{a \in Set_{W'}} \text{GetProb}(a)$$

Log-linear model for chunk alignment (cont.)

- Distance distortion (F-DIST)
  - Predict the unknown from the known
  - guess the probability of ambiguous chunk alignment using its distortion from the chunk alignment without ambiguity (anchor chunk)



Log-linear model for chunk alignment (cont.)



i (j), position of Chinese (English) word

$$\text{LocDist}(i,j)=\min(|\text{Slope}_L-1|,|\text{Slope}_R-1|)$$

$$\text{Slope}_L=(j-j_L)/(i-i_L),\quad \text{Slope}_R=(j_R-j)/(i_R-i)$$

$$(i_L, j_L) = \arg\max_{(cp_{i'}, i', 1, ep_{j'}, j', l_{ep}) \in Set_{PA<i}} i' \qquad (i_R, j_R) = \arg\max_{(cp_{i'}, i', 1, ep_{j'}, j', l_{ep}) \in Set_{PA>i}} i'$$

the nearest anchor chunk on the left

LocDist(2, 2) = 0 $\text{Slope}_L=(2-1)/(2-1)=1$; $\text{Slope}_R=(5-2)/(3-2)=3$

LocDist(2, 6) = 2 $\text{Slope}_L=(6-1)/(2-1)=5$; $\text{Slope}_R=(5-6/3-2)=-1$

Log-linear model for chunk alignment (cont.)

- the less LocDist(i,j) is, the more probable (i,j) is

$$F - DIST = \frac{1}{LocDist(cp_j, ep_i^{i+n})}$$

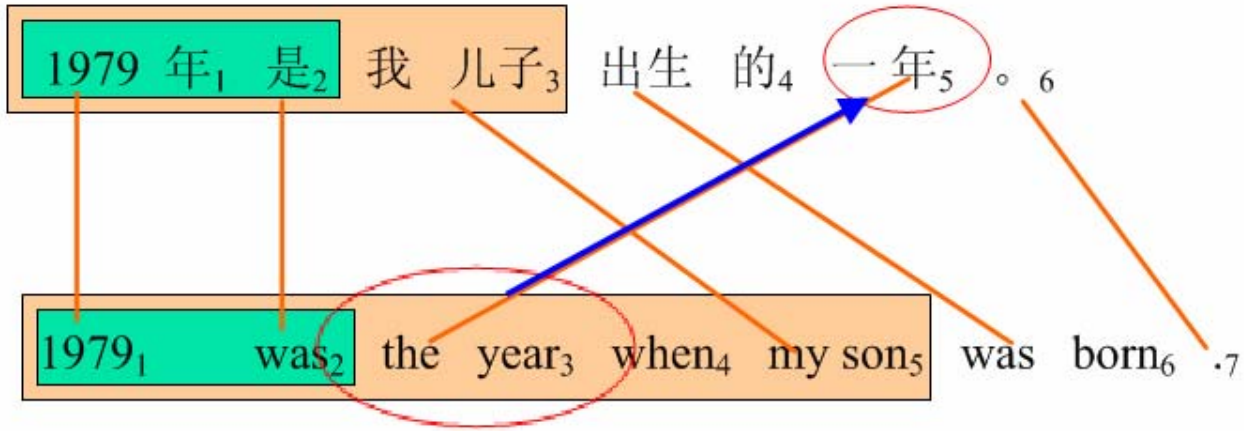
- Marker translation table (F-MARK)
  - Matrix of translation probability from on marker to another

Log-linear model for chunk alignment (cont.)

- Feature weighting
  - YASMET toolkit
- Formula for candidate scoring

$$Score(< cp_j, ep_i^{i+n} >) = \lambda_1 * F - WALI(< cp_j, ep_i^{i+n} >) +$$
$$\lambda_2 * F - DIST(< cp_j, ep_i^{i+n} >) +$$
$$\lambda_3 * F - MARK(< cp_j, ep_i^{i+n} >)$$

Extracting n:n chunk alignment from 1:n chunk alignment

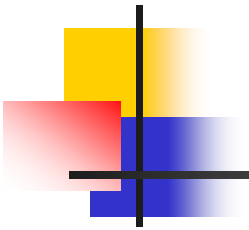


- n:n chunks

  1979年 是：1979 was

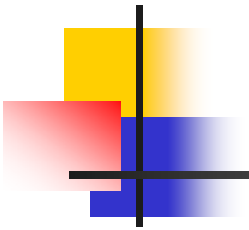
  我 儿子 出生：my son was born

  ……



Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work



Experiments & analysis

- Word alignment
- BaseNP identification and alignment
- Chunk alignment

Experiments & analysis--- Anchor word alignment

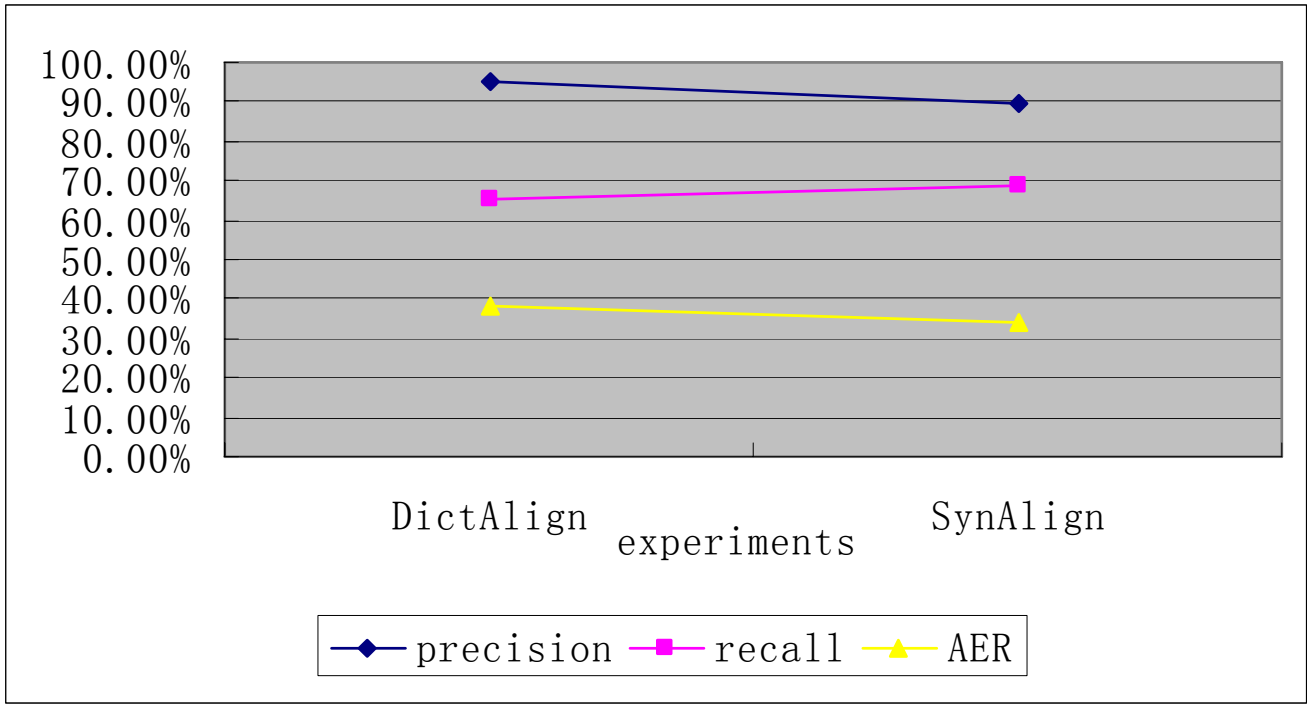
- Evaluation metric：in answer set, sure word alignment is SW, possible word alignment is PW. Word alignment of our system is AW.

$$precision = \frac{|AW \cap PW|}{|AW|} \qquad recall = \frac{|AW \cap SW|}{|SW|}$$

$$AER(SW, PW; AW) = 1 - \frac{|AW \cap SW| + |AW \cap PW|}{|AW| + |SW|}$$

Experiments & analysis--- Anchor word alignment

experiments	precision	recall	ASR
DictAlign	94.97%	65.27%	0.3799
SynAlign	89.69%	69.01%	0.3416



Experiments & analysis---BaseNP identification and alignment

- Evaluation
  - In answer set the number of Chinese BaseNP is SNP, the number of Chinese BaseNP which can be aligned to English BaseNP is NNP. The number of Chinese BaseNP identified by our system is MNP, the number of correct BaseNP is CNP, that can be correctly aligned is ANP.
  - BaseNP identification

$$precision = \frac{CNP}{MNP} \qquad recall = \frac{CNP}{SNP}$$

  - BaseNP alignment

$$precision = \frac{ANP}{MNP} \qquad recall = \frac{ANP}{NNP}$$

Experiments & analysis---BaseNP identification and alignment

- results

Experiments	precision	recall	FB1
BaseNP identification	92.77%	93.15%	92.96
BaseNP alignment	89.26%	73.60%	80.68

Experiments & analysis--- chunk alignment

- Evaluation: extract all the aligned fragments in 200 sentence pair corpus as the answer set. The set of sure alignment is SP, possible alignment is PP.
- The set of aligned chunks in our system is AP

$$precision = \frac{|AP \cap PP|}{|AP|} \qquad recall = \frac{|AP \cap SP|}{|SP|}$$

Experiments & analysis--- chunk alignment

- Chunk identification methods
  - Marker chunks（MARKER_CHUNK）
  - Hybrid chunks（HYBRID_CHUNK）
- Experiment scheme
  - 1:n chunk alignment with heuristics（SM_AC）
  - 1:n chunk alignment plus log-linear model ranking（SM_AC+ME）
  - n:n chunk alignment: extension of SM_AC（MM_AC）
  - n:n chunk alignment: extention of SM_AC+ME（MM_AC+ME）

Experiments & analysis---chunk alignment

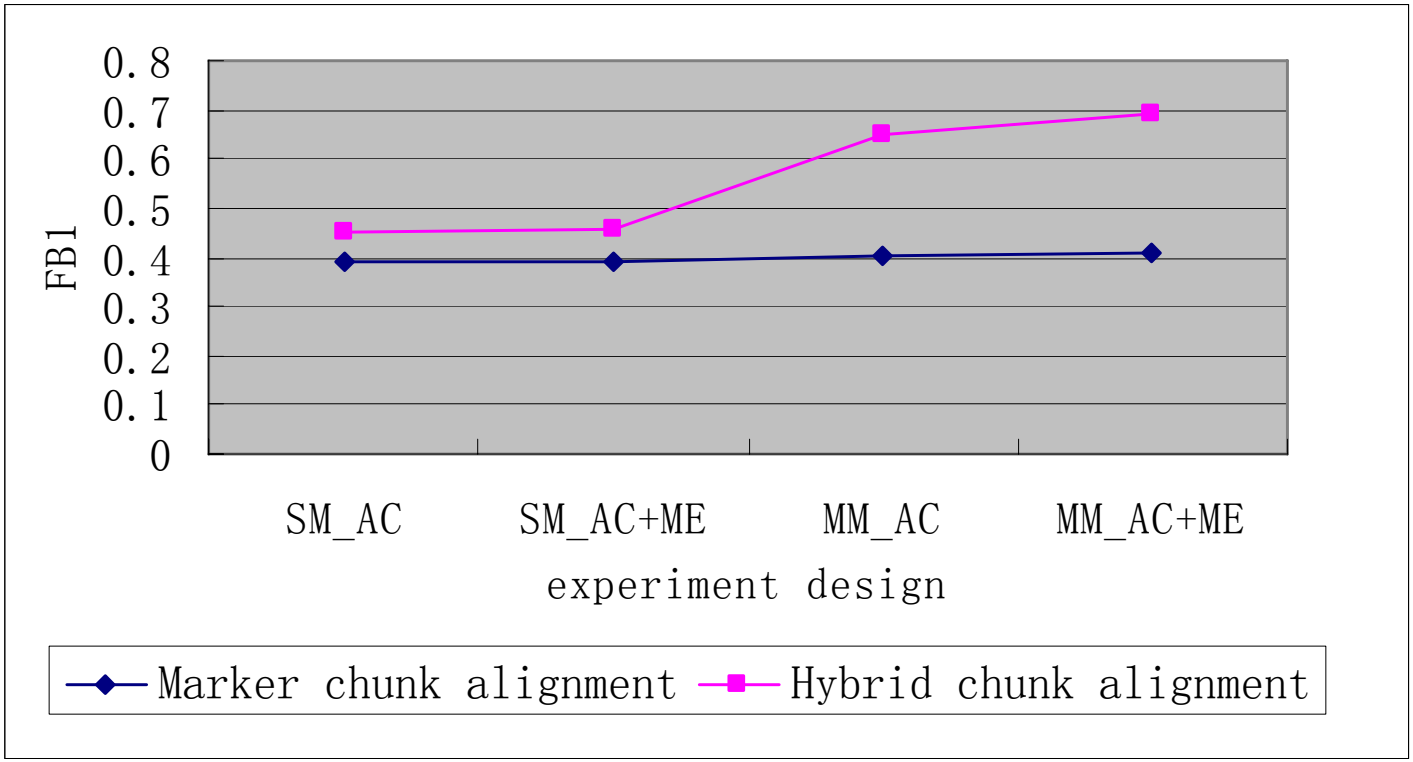
- Marker chunk alignment

Experiments	precision	recall	FB1
SM_AC	86.84%	25.40%	39.30
SM_AC+ME	78.57%	26.12%	39.21
MM_AC	85%	26.19%	40.04
MM_AC+ME	76.60%	27.78%	40.77

- Hybrid chunk alignment

Experiments	precision	recall	FB1
SM_AC	88.64%	30.16%	45
SM_AC+ME	83.67%	31.75%	46.05
MM_AC	90.59%	50.79%	65.09
MM_AC+ME	89.69%	56.35%	69.21

Experiments & analysis---chunk alignment



Experiments & analysis---chunk alignment

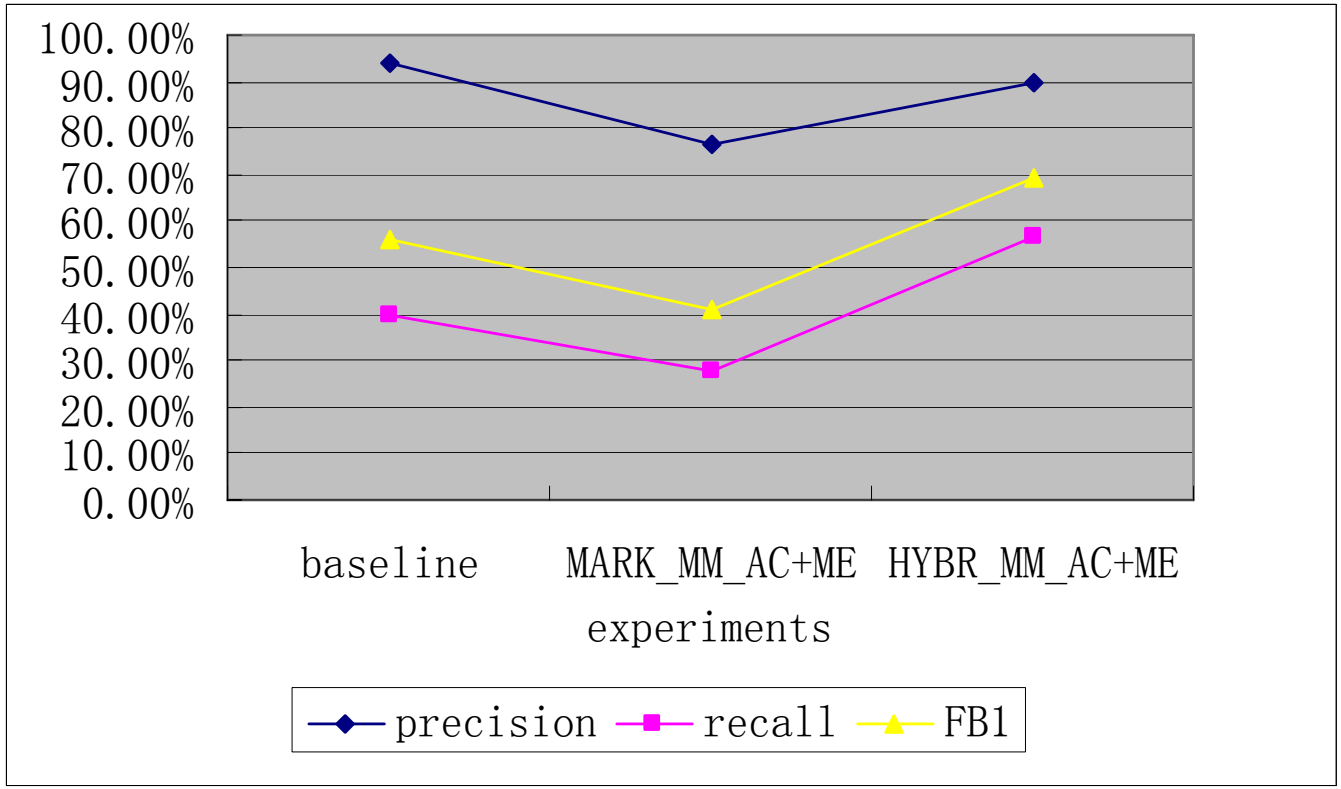
- Chunk alignment based on Hybrid chunk identification outperforms that based on Marker chunks.
- Extraction of n:n chunk alignments can improve recall significantly while the precision is slightly lower.
- In 1:n chunk alignment, ranking module makes slight contribution. While in n:n alignment, ranking module contributes greatly to recall.

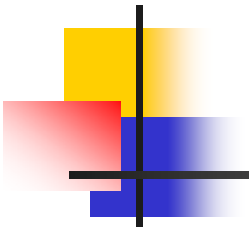
Experiments & analysis---chunk alignment

- Comparison of chunk alignment with different chunk identification methods

experiments	precision	recall	FB1
SMT chunks	94.03%	39.68%	55.81
MARK_MM_AC+ME	76.60%	27.78%	40.77
HYBR_MM_AC+ME	89.69%	56.35%	69.21

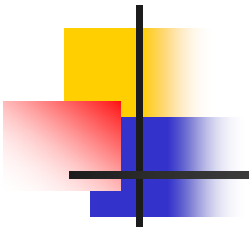
Experiments & analysis---chunk alignment





Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work



Future work

- Word alignment
  - Improve recall
  - Statistical methods
- Evaluation metric for chunk alignment
- Relationship between different chunk alignment

Thanks!